UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Segmentación de Audio en Vídeos de Redes Sociales

Máster Universitario en Ingeniería de Telecomunicación

Autor: SORIANO MORANCHO, Gonzalo
Tutor: GONZÁLEZ RODRÍGUEZ, Joaquín

FECHA: Febrero, 2018

Segmentación de Audio en Vídeos de Redes Sociales

AUTOR: Gonzalo Soriano Morancho
TUTOR: Joaquín González Rodríguez

Audias- Audio, Data Intelligence and Speech
Dpto. Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Febrero de 2018

Agradecimientos

Muchas gracias a mi tutor Joaquín, porque tras guiarme en mi Trabajo Fin de Grado, también ha querido ayudarme a la consecución de este Trabajo Fin de Máster.

A los compañeros del grupo Audias, especialmente a Álvaro y Alicia, por sus continuos ofrecimientos de ayuda y su disponibilidad total para llevarlos a cabo de verdad.

A mi familia y amigos, porque siempre me han ayudado a escoger los mejores caminos en mi vida y, ante dificultades, nunca he dudado de su total disposición a afrontarlas juntos.

A mi padre y mi abuela.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 Introducción

1.1 Motivación

¿Qué desafíos técnicos presentan las redes sociales para el procesado de Audio?

Las redes sociales han cambiado para siempre la forma de relacionarnos con la gente, desde comunicarnos en directo con gente al otro lado del mundo, hasta difundir contenido multimedia creado por nosotros mismos para que sea consumido por todo aquel que lo desee.

Este cambio, propiciado por plataformas como *Youtube*, hace que el contenido audiovisual que consumimos sea cada vez menos controlado, más diverso y, sobre todo, más abundante, cada minuto se suben 300 horas de vídeo sólo a *Youtube*.

Organizar tal cantidad de contenido supone multitud de retos en diversas áreas de conocimiento: procesado de vídeo, procesado de audio, utilización de Big Data…

Pero ¿cuál es el resultado de utilizar un sistema de detector de emociones en audio musical? ¿y en un sistema de detección de idioma? ¿qué ocurre si sólo hay ruidos de fondo? Estos sistemas precisan de voz para poder funcionar correctamente, la introducción de otros elementos puede hacer que se lancen resultados que no se corresponden con ninguna realidad.

Aquí está la principal motivación de este Trabajo Fin de Máster, tratar de proponer una solución a un problema -el de segmentar audio- que parecía resuelto, pero que ha vuelto a manifestar su importancia gracias a las redes sociales.

Tal es la importancia que ha adquirido esta tarea, que el *National Institute of Standards and Technology* -en adelante, NIST-, ha lanzado, por primera vez en su historia, una evaluación específica sobre segmentación de audio con corpus muy desafiantes en este aspecto.

La gran importancia de los sistemas de *Voice Activity Detection* y su evaluación por el NIST en OpenSAT 2017, crearon un gran interés dentro del grupo Audias-ATVS para evaluar el rendimiento de sus sistemas dentro de un marco común que permita la comparativa entre cada una de las distintas propuestas con grupos de investigación punteros en este tipo de tecnologías.

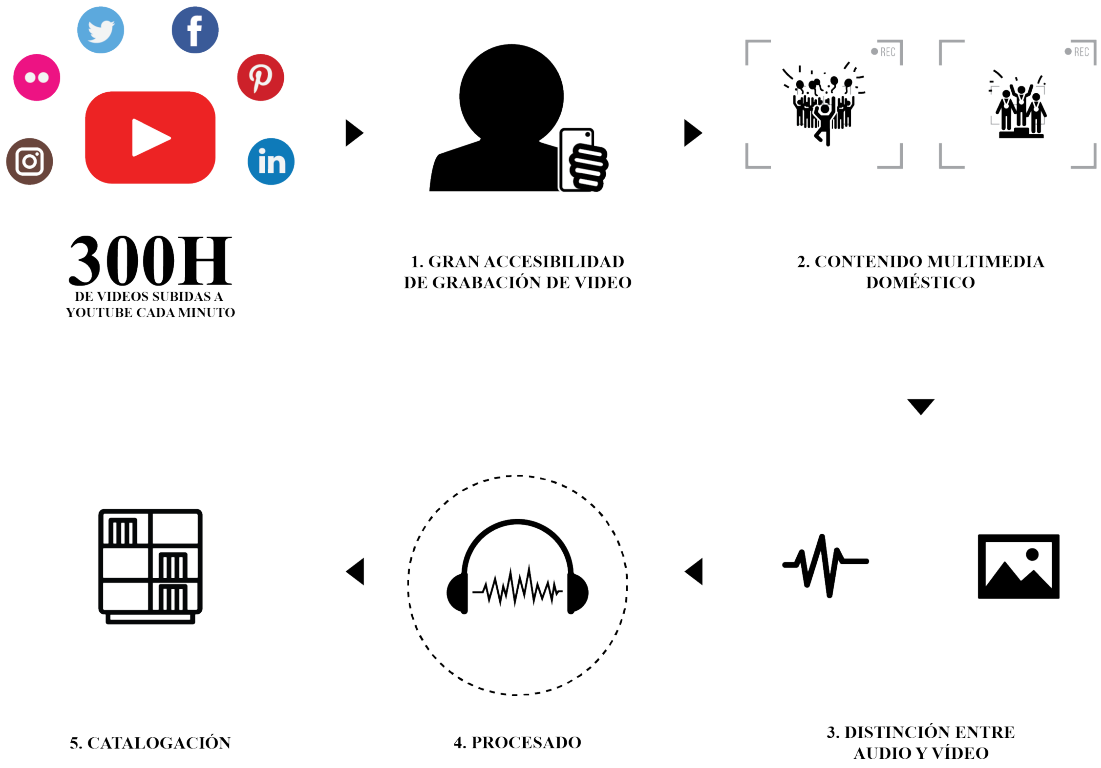

Figura 1-1: Gráfico explicativo de la necesidad de un sistema de segmentación de audio.

1.2 Objetivos

El objetivo fundamental de este proyecto consiste en realizar un estudio del rendimiento de distintas alternativas de sistemas para este propósito y proponer mejoras a los mismos.

Para ello, en primer lugar, habrá que hacer un estudio de las herramientas del estado del arte disponibles, entendiendo su funcionamiento, analizando sus fortalezas y debilidades. Este paso se hace indispensable a la hora de proponer mejoras a los distintos sistemas.

De manera secundaria, se ha dado el primer paso en la creación de un corpus de trabajo. Para ello, se ha partido de la base de datos existente VAST y se ha procedido a un etiquetado a tres niveles de cinco horas de audio. La gran característica de esta base de datos es su gran diversidad, dado que se trata de audios obtenidos de redes sociales, en las que cada usuario puede subir casi cualquier tipo de contenido.

Es importante tener en cuenta que los sistemas que analizaremos se han presentado a la evaluación OpenSAT del NIST. Por ello, nuestra métrica de rendimiento será, desde el primer momento, la misma con la que estos sistemas eran evaluados.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Capítulo 1. Introducción**

En el primer capítulo de este TFM se explican las motivaciones y objetivos perseguidos, además de realizar una breve explicación de la estructura del mismo.

- **Capítulo 2. Estado del arte**

Este capítulo está dedicado a un estudio del estado actual de los sistemas y tecnologías que se estudiarán. Se dará una introducción sobre todos los sistemas que se utilizarán y se verá cuál es su base de funcionamiento.

- **Capítulo 3. Entorno experimental**

Se hace una descripción de la metodología seguida para crear el set de utilidades que se utilizarán para hacer las pruebas.

- **Capítulo 4. Pruebas y resultados**

En este capítulo se realizan todas las pruebas con las herramientas bajo estudio y se analizan los resultados de las mismas.

- **Capítulo 5. Conclusiones y trabajo futuro**

Se recopila todo lo realizado y estudiado a lo largo del TFM y se proponen posibles líneas de actuación futuras.

2 Estado del arte

El sistema presentado por Audias-ATVS a la evaluación OpenSAT está formado por la fusión de dos sistemas: un *speech activity detector* [1] y una *deep neural network*. Como primer paso, deberemos entender el funcionamiento de los mismos con el objetivo de proponer mejoras a su diseño más adelante.

Los *voice activity detectors* clásicos basaban su funcionamiento en características directas de la señal de voz, normalmente la energía a corto plazo de la señal o su tasa de cruces por cero. Sin embargo, estas características se degradan con mucha facilidad en presencia de ruido. Por esta razón se han propuesto a lo largo de los años multitud de características más robustas: características basadas en la función de autocorrelación, características basadas en espectro, Mel Frequency Cepstral Coefficients…

2.1 Speech Activity Detector + Music Activity Detector

Este sistema, descrito en [1] y [12], busca hacer un estudio de la correlación de la señal de audio para hacer la segmentación.

La entrada tipo que recibe un sistema como este podría ser similar al siguiente ejemplo:

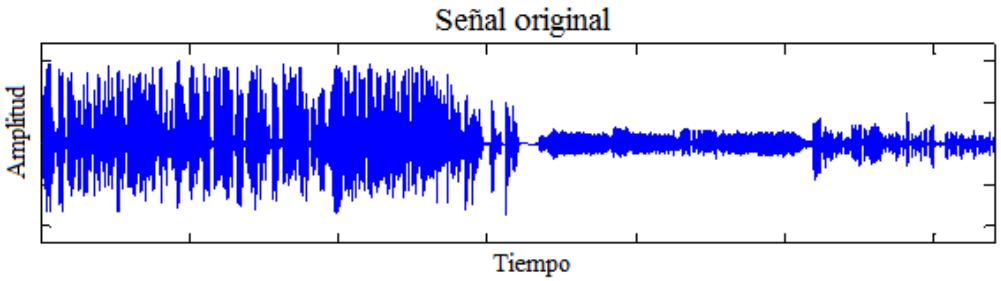

Figura 2-1: Señal cualquiera de entrada a un sistema de detección de voz, extraída de [1].

Tal y como se puede observar, puede ser que la entrada al sistema presente niveles muy dispares en amplitud a lo largo del tiempo. Ya que este sistema no se basará en un estudio de la energía de la señal, será interesante para el mismo que las características intrínsecas de la señal se independicen al máximo de ella. Es por esta razón que la primera etapa del sistema consiste en una normalización de la ganancia del sistema a lo largo del tiempo.

2.1.1 Normalización de la ganancia

Debido a la naturaleza tan diversa de nuestra base de datos de estudio, la energía de la señal de audio a lo largo de la misma puede llegar a ser muy notable. Incluso, puede existir gran variabilidad en el propio archivo, por ejemplo, un archivo que mezcle locuciones en primer plano con locuciones en segundo plano. Es por estas razones que la primera etapa del sistema consiste en una normalización de la ganancia de la señal.

Tal y como está expuesto en [1] el esquema que se ha seguido para la normalización es la utilización de una señal que multiplica a la original. Dicha señal está compuesta por una serie de ventanas triangulares de cinco segundos de duración con un solapamiento entre ellas del 50%. Debido a que cada una de estas ventanas tiene una ganancia distinta, se obtiene como resultado de la multiplicación una señal más uniforme.

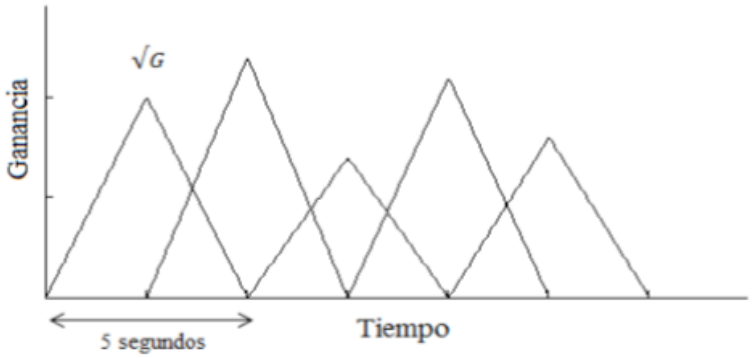

Figura 2-2: Señal compuesta por ventanas triangulares que multiplica a la señal original con el objetivo de normalizar su ganancia, propuesta en [1].

Ahora se ha de mirar la forma en la que se calcula la ganancia de cada ventana triangular, para ello en [1] se utiliza el siguiente proceso:

- Enventanado de cada ventana utilizando ventanas de 30 ms con un solapamiento entre ellas de 10 ms.
- Cálculo de la energía y su logaritmo para cada ventana:

$$E(k) = \frac{1}{N}\sum x^2$$

$$E_{log} = 10 * \log(E(k))$$

- Aplicación de un umbral de decisión para el valor E_{log}, de tal forma que:
  - Si su valor es mayor que 20 dB, se considerará silencio todo lo que sea menor que el 30% de la misma.
  - Si su valor es menor que 20 dB, se considerará que todo es señal de no silencio.
- Se calcula la energía logarítmica de la parte de la señal que no es silencio x_{nosil}.

$$E_{nosil} = \frac{1}{N_{muestrasnosil}} \sum E(k)_{nosil}$$

- Por último, se calcula la ganancia -G-:

$$G = \sqrt{\frac{1}{E_{nosil}}}$$

Un ejemplo de utilización de esta corrección de ganancia se puede observar en la siguiente figura:

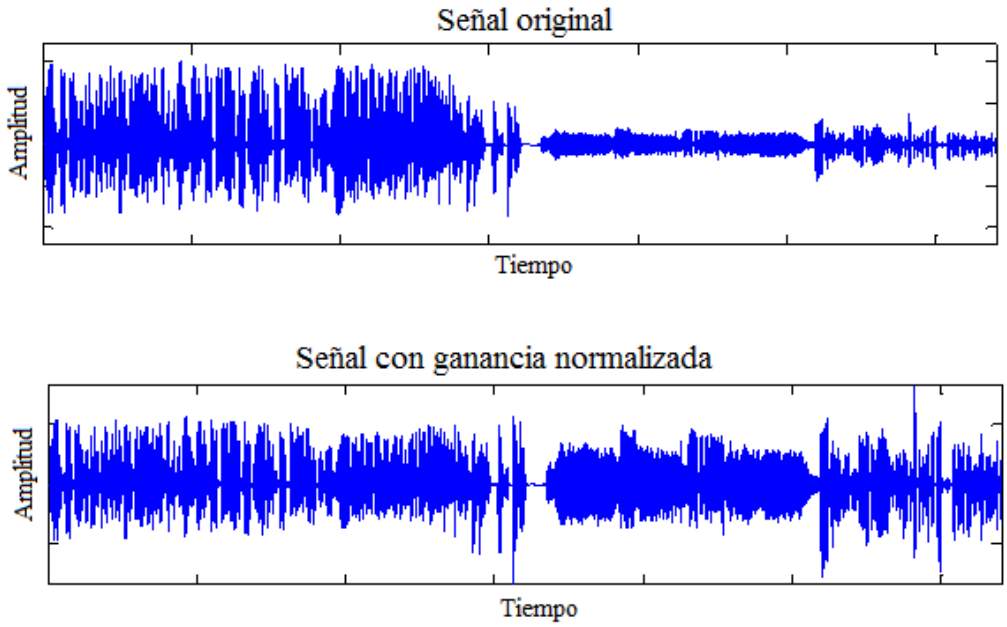

Figura 2-3: Ejemplo de señal original y su normalización de ganancia en una señal de audio de 25 segundos de duración, extraída de [1].

Donde podemos ver que las partes que tenían menos energía han sido igualadas en gran medida a la primera parte del audio, que tenía un valor de amplitud mucho más grande.

2.1.2 Enventanado de la señal

Una vez tenemos una señal cuyos niveles de amplitud a lo largo del tiempo están normalizados, se pasa a estudiar en detalle la señal de entrada. Para ello, el primer paso ha de ser enventanar la señal, este paso es indispensable en audio y se realiza debido a las características intrínsecas de la señal de voz. En este caso se utilizan ventanas de 30 ms con un solapamiento temporal entre ellas de 10 ms. De las alternativas existentes, el sistema [1] se decanta por una ventana tipo Hamming definida por:

$$w(n) = a_0 - a_1 * \cos\frac{2\pi n}{N-1}$$

donde a_0 = 0,5383 y a_1 = 0,4616 y donde w(n) tomará ese valor siempre que nos encontremos en N-1 ≥ n ≥ 0.

2.1.3 Representación del Melgrama

Un Melgrama no es otra cosa que una adaptación de un espectrograma a escala de frecuencias de Mel. En nuestro caso, debido a que deseamos identificar trayectorias armónicas en la señal de voz -paralelas en el espectrograma-, esta normalización hace que las trayectorias no tiendan a divergir, tal y como podemos ver en la siguiente figura:

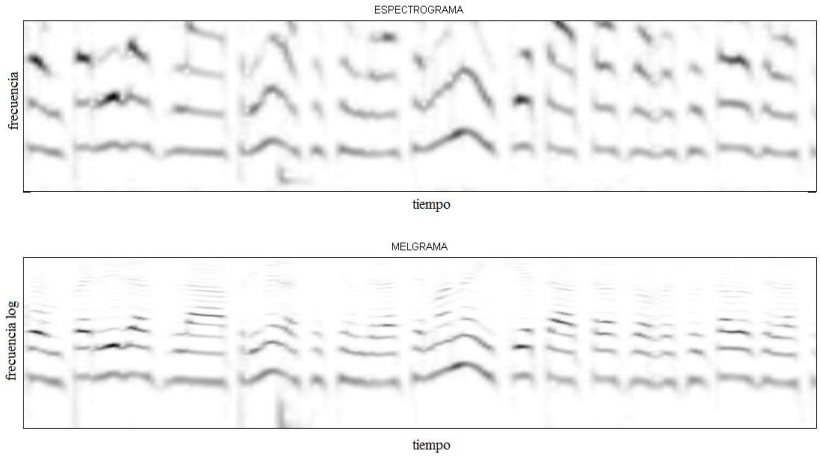

Figura 2-4: Espectrograma y Melgrama de un segmento de audio, extraída de [1].

En este caso se han utilizado 6 octavas de 40 bins cada una, variando de la siguiente manera:

- Octava 1: 50 Hz – 100 Hz.
- Octava 2: 100 Hz – 200 Hz.
- Octava 3: 200 Hz – 400 Hz.
- Octava 4: 400 Hz – 800 Hz.
- Octava 5: 800 Hz – 1600 Hz.
- Octava 6: 1600 Hz – 3200 Hz.

Así pues, añadiendo las octavas 5 y 6 se añade resolución para el estudio de las señales para voz de mujeres y música. Estas octavas se utilizaban como frecuencias límite para 6 ventanas triangulares en frecuencia, llevando a cabo la mencionada normalización a Escala Mel.

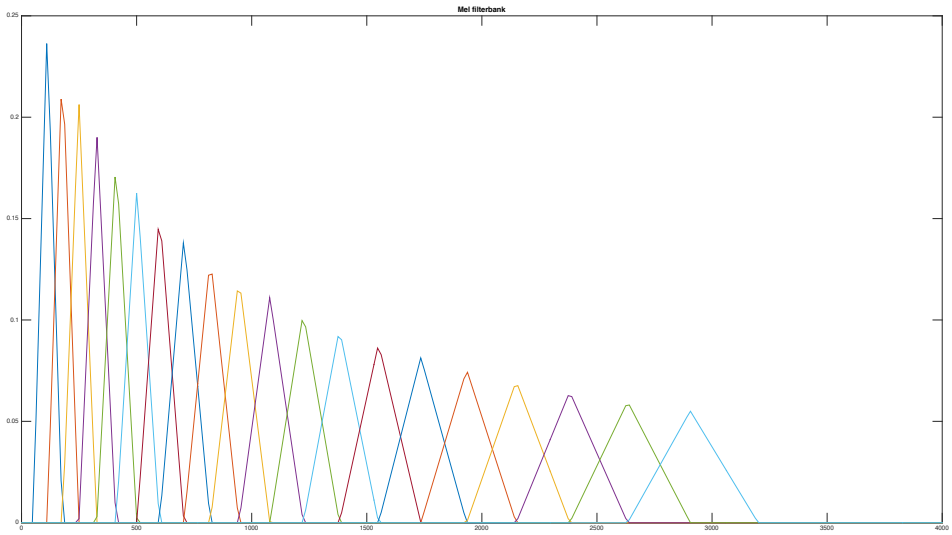

Figura 2-5: Banco de filtros en Escala Mel utilizado.

2.1.4 Cálculo de las correlaciones

El resultado de la normalización previa en busca de trayectorias paralelas en el Melgrama, buscaba simplificar y mejorar este paso. El hecho del paralelismo en el Melgrama se traduce en una alta correlación en la señal cuando comparamos los espectros en dos tramas de audio de ventanas lindantes en el tiempo, y es en este hecho en el que se busca hacer la distinción entre voz y música.

Para el cálculo de esa, presumiblemente alta, correlación se deberá calcular la correlación cruzada de cada una de las columnas de la matriz del Melgrama con sus columnas cercanas -4 en este caso-. Además, con el fin de corregir pequeñas desviaciones temporales, se tendrán en cuenta 20 muestras adicionales -10 hacia arriba y 10 hacia abajo-.

El resultado de estos cálculos será una matriz del que se obtendrá el máximo por columna para hallar el vector r_{xcorr}.

$$r_{xcorr}(X_t, X_{t+offset}) = \max(R_{X_t, X_{t+offset}}(l))$$

donde $l \in [-l_{max}, l_{max}]$ marca el desplazamiento en el eje de frecuencias. Se define también el caso particular para el cual $l = 0$ quedando:

$$r(X_t, X_{t+offset}) = R_{X_t, X_{t+offset}}(0)$$

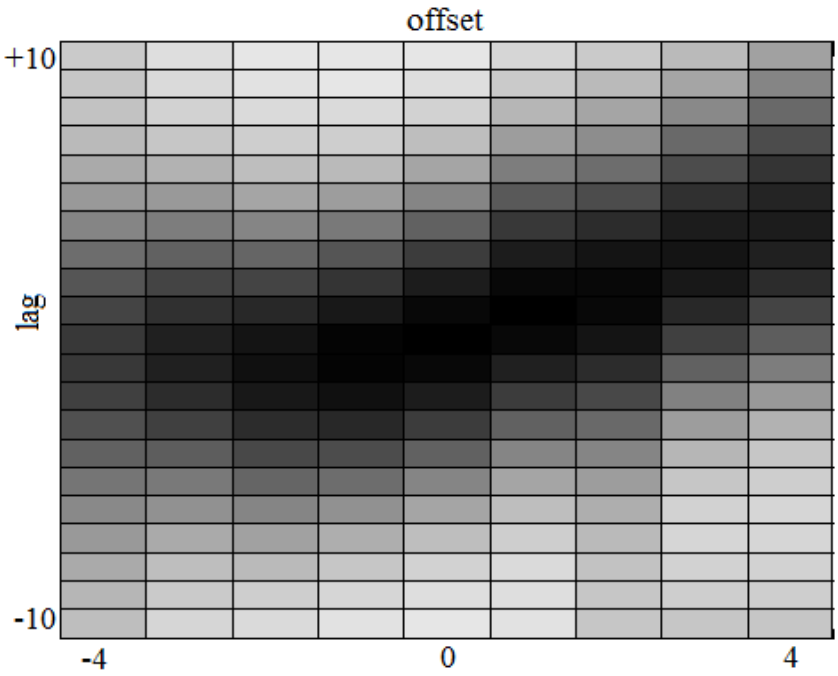

Figura 2-6: Matriz de correlación calculada a partir de la correlación de un vector de la matriz de melgrama, extraída de [1].

Llegado a este punto se podría obtener un indicador de presencia de voz, definiendo un valor $r_{xcorr} - r$. Este valor, para señales de música, será máximo para $l = 0$ y, por tanto, el valor de $r_{xcorr} - r = R(0) - r = r - r = 0$. Por otra parte, para señales que presenten armónicos no tendrán el máximo en $l=0$ y, por tanto, $r_{xcorr} - r$ será mayor que cero.

2.1.5 Estudio de las trayectorias y scores

El cálculo obtenido hasta ahora proporciona buenos resultados, sin embargo, se pueden observar picos espurios en zonas de música -donde el valor debería ser bajo-, y zonas anormalmente bajas en zonas de voz -donde el valor de correlación debería ser alto. Tal y como se puede observar en la siguiente figura:

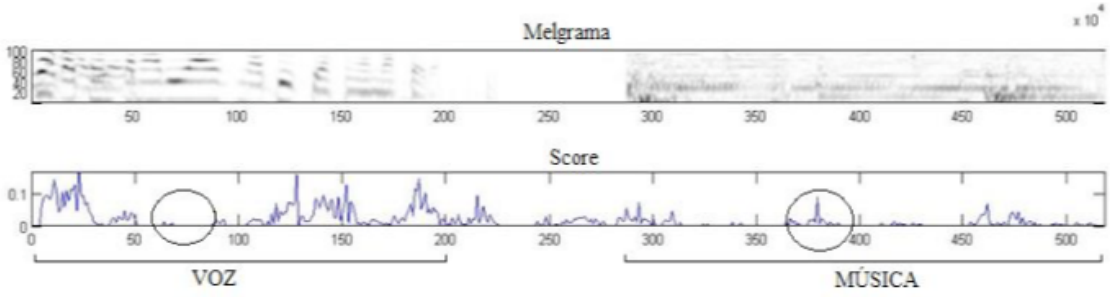

Figura 2-7: Resalte de fallos en la correlación, segmentos de voz con valores muy bajos y segmentos de música con picos, extraída de [1].

Para la corrección de estos fallos se realiza un estudio de la trayectoria, apreciándose la continuidad antes mencionada en los tramos con voz.

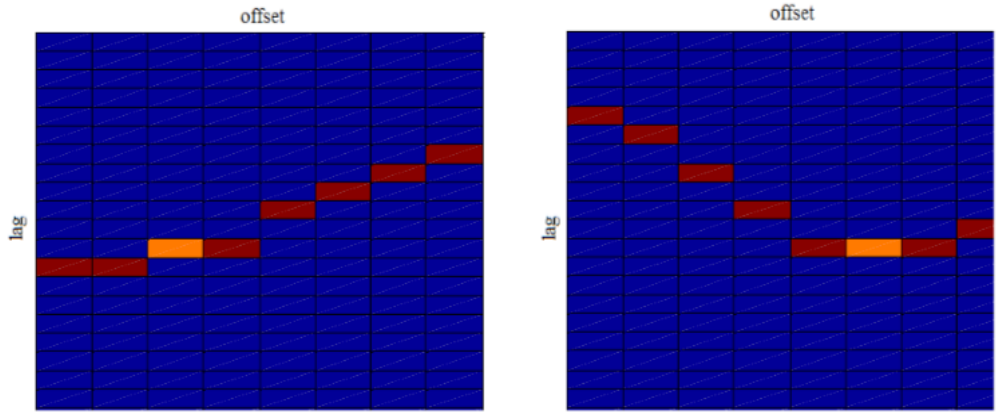

Figura 2-8: Trayectorias de dos tramos de voz, extraída de [1].

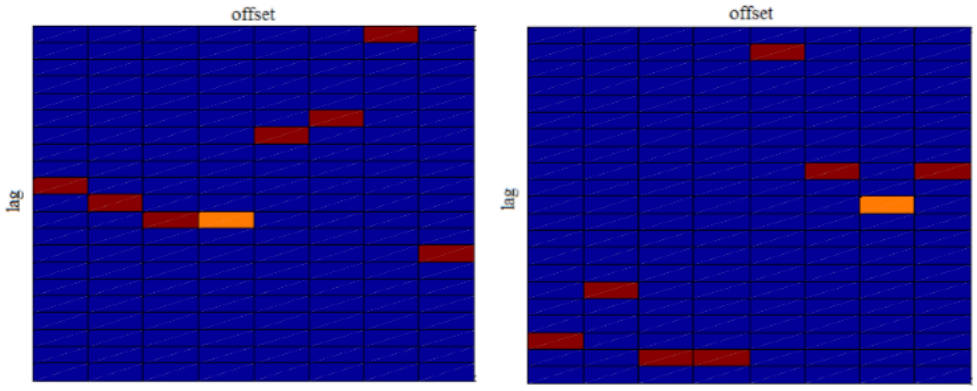

Figura 2-9: Trayectorias para dos tramos de no voz, extraída de [1].

Estos resultados hacen ver la bondad de un estudio de la matriz de correlaciones que pueda permitir eliminar fallos como los vistos en la figura 2-6. La trayectoria que se estudiará será el vector que antes hemos denominado como r_{xcorr}.

La determinación de un score para la trayectoria se realiza siguiendo estos cuatro pasos:

1. Descartar los valores que son menores que el 30% de la energía logarítmica antes hallada.
2. Descartar aquellos valores que tienen una variación con respecto al anterior de más del 20%.
3. Cálculo de un score tal que:

$$score = \frac{\sum r_{xcorr}(l = punto\ de\ estudio)}{N_{puntos}}$$

teniendo en cuenta que, si N_{puntos} es menor que cuatro, a la hora del cálculo del score se pondrá $N_{puntos} = 4$.

4. Cálculo del score en escala logarítmica.

2.1.6 Detector de pitch

El pitch, o frecuencia promedio a la que vibran las cuerdas vocales, es un parámetro que se añade con el objetivo de mejorar la detección de voz. Esa mejora no se producirá tanto en la medida en sí del pitch, sino que lo que realmente ayudará será la medida que en [1] se denomina "fiabilidad del pitch".

La estimación del pitch mediante correlación se realiza calculando la autocorrelación de cada tramo de señal y luego se calcula un máximo local -se excluye el máximo global-.

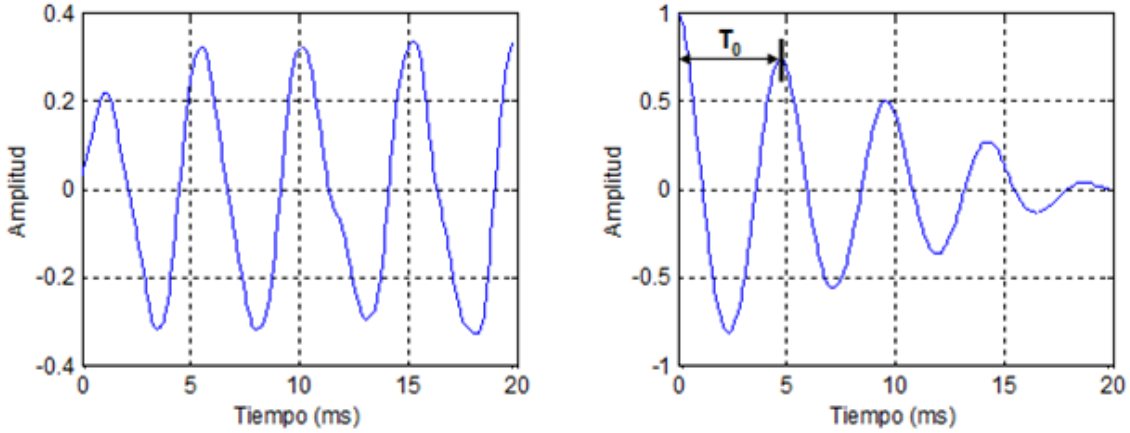

Figura 2-10: Extracción del pitch por correlación, extraída de [1].

Una vez obtenido el valor T_0 el pitch será simplemente $f = 1/T_0$. Mientras que el parámetro "fiabilidad del pitch" se corresponde con el valor del primer pico distinto de cero.

2.1.7 Corrección por pitch constante

Una vez analizados los resultados, se observaron niveles de pitch anormalmente altos -llegando a 900 Hz-, reproduciendo el audio se observó que esta circunstancia se correspondía con la aparición de los tonos usados en radios para la marca de las horas.

Es por ello, que, directamente, se eliminan esos valores anormalmente altos que podrían influir negativamente en el algoritmo. Dado que el pitch más alto correspondiente a voz se corresponderá a una mujer y como máximo se llega a 400 Hz, este sistema coloca un umbral a 500 Hz, a partir del cual ya no se considerará voz.

Observando los segmentos donde había voz, se ve que, en la mayoría de ellos, la variación del pitch es prácticamente nula.

Siguiendo este razonamiento se lleva a cabo una búsqueda de aquellos segmentos en los que el pitch varíe menos del 10% y, además, la fiabilidad del pitch esté por encima de un umbral. En caso de cumplirse ambas condiciones, se ignorará el resultado del estudio de trayectorias y se marcará el segmento como voz.

2.1.8 Score final del SAD

Este sistema realiza la decisión siguiendo los siguientes pasos:

1. Cálculo del histograma del score instantáneo.
2. Cálculo de la suma acumulada de los valores del histograma.

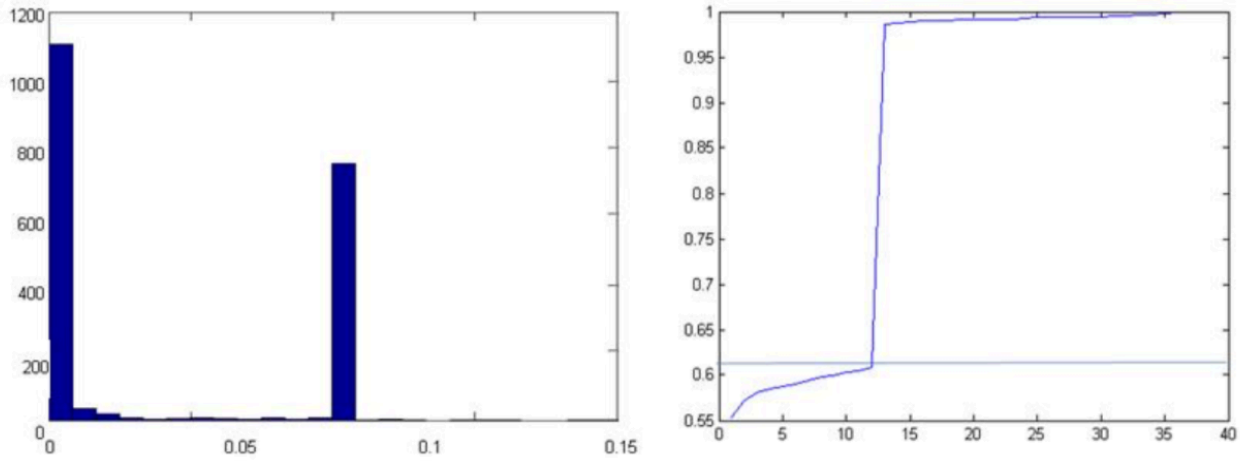

Figura 2-11: Histograma de score (izquierda) y su suma acumulada (derecha) para el sistema de Speech Activity Detection, extraída de [1].

3. Cálculo del valor que se corresponde con el 10% de esa suma acumulada.
4. Comprobación del valor del histograma con el que se corresponde ese valor.
5. Decisión, si el score está por encima del valor obtenido en el paso 4, será voz y si no, no voz.

Sin embargo, después de este proceso, se obtienen tramas espurias en medio de una locución indicando que no son voz. Es por esto por lo que el sistema propone tomar en consideración el entorno de cada trama, de tal forma que se puedan eliminar aquellos espurios en medio de una locución. Todo esto se puede ver en la siguiente figura, observando la segunda gráfica y la tercera, como han desaparecido los valores intermedios.

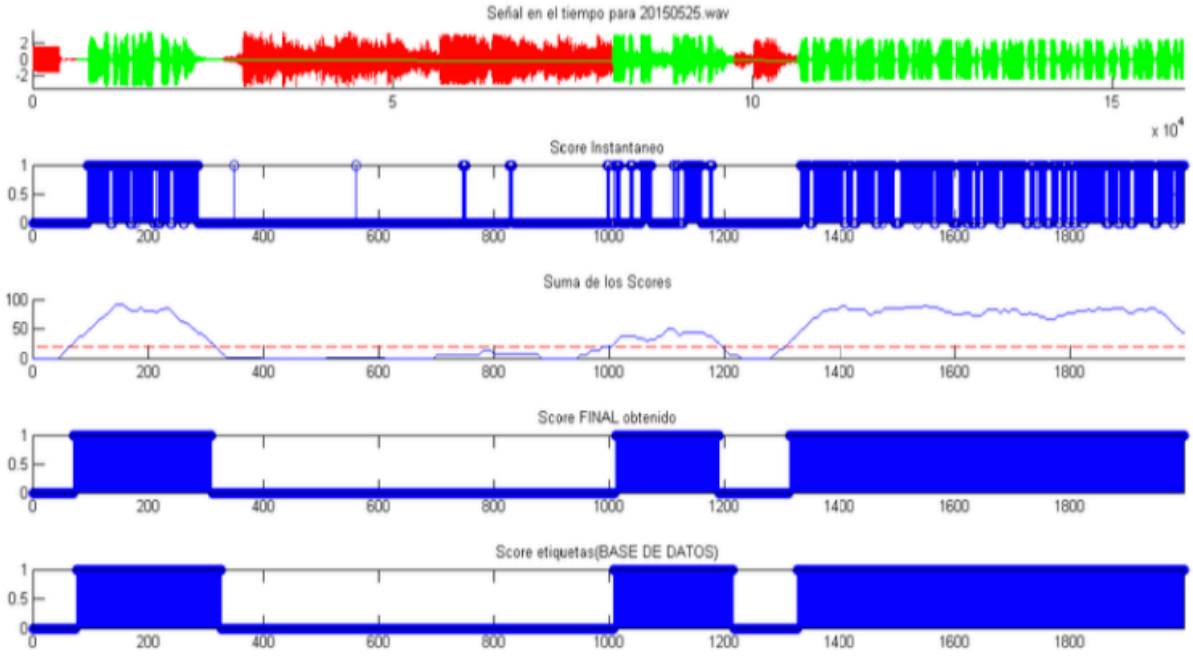

Figura 2-12: Score final del sistema tras hacer un suavizado mediante la observación del entorno de cada trama en el Speech Activity Detection, extraída de [1].

Finalmente, se hace una transformación de los resultados en base a la confianza y al score obtenido de la forma:

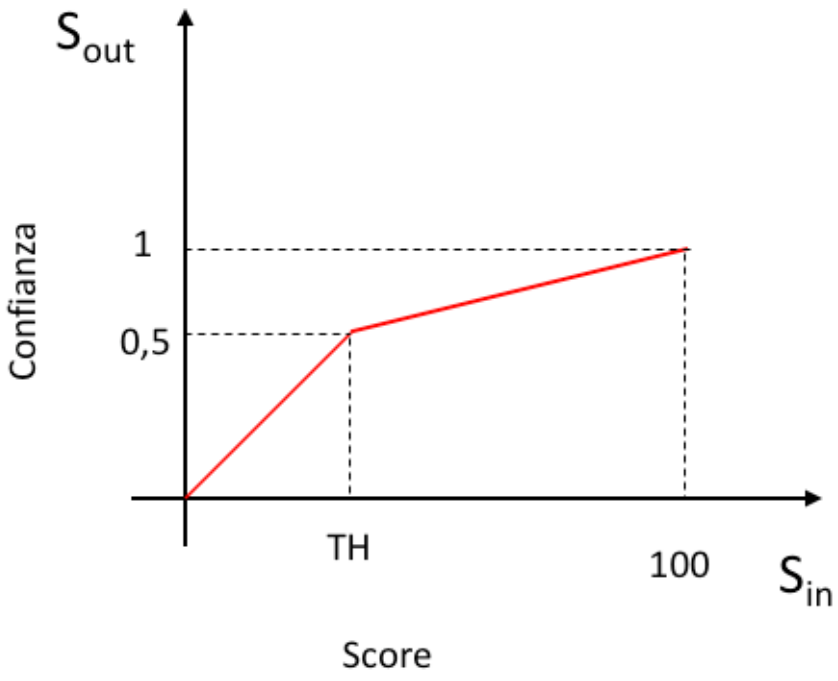

Figura 2-13: Transformación de Score de acuerdo con la confianza y score obtenidos.

2.1.9 Music Activity Detector

Este sistema, explicado en [12], tiene el condicionante de cómo se defina musicalidad. Éste es un concepto subjetivo que, sin embargo, en [12] es definido de tal forma que se fijan dos condiciones para que un segmento de audio sea considerado musical, dichas condiciones son:

1. Se mantiene un pulso rítmico constante -o casi- y dentro de un margen de velocidad o tempo.
2. Hay notas musicales en el espectro que son favorecidas frente a otras.

La primera condición será evaluada por un detector de ritmo, mientras que la segunda se evaluará en un detector de armonía. Ambos resultados serán luego combinados de tal forma que el sistema termine dando una decisión sobre la presencia o no de música.

2.1.9.1 Detector de ritmo

El detector de ritmo se basará en un cálculo de correlaciones ya visto en [1] y explicado en este TFM en el apartado 2.1.4. Para un segmento de música, este sistema obtendría unas gráficas como las siguientes:

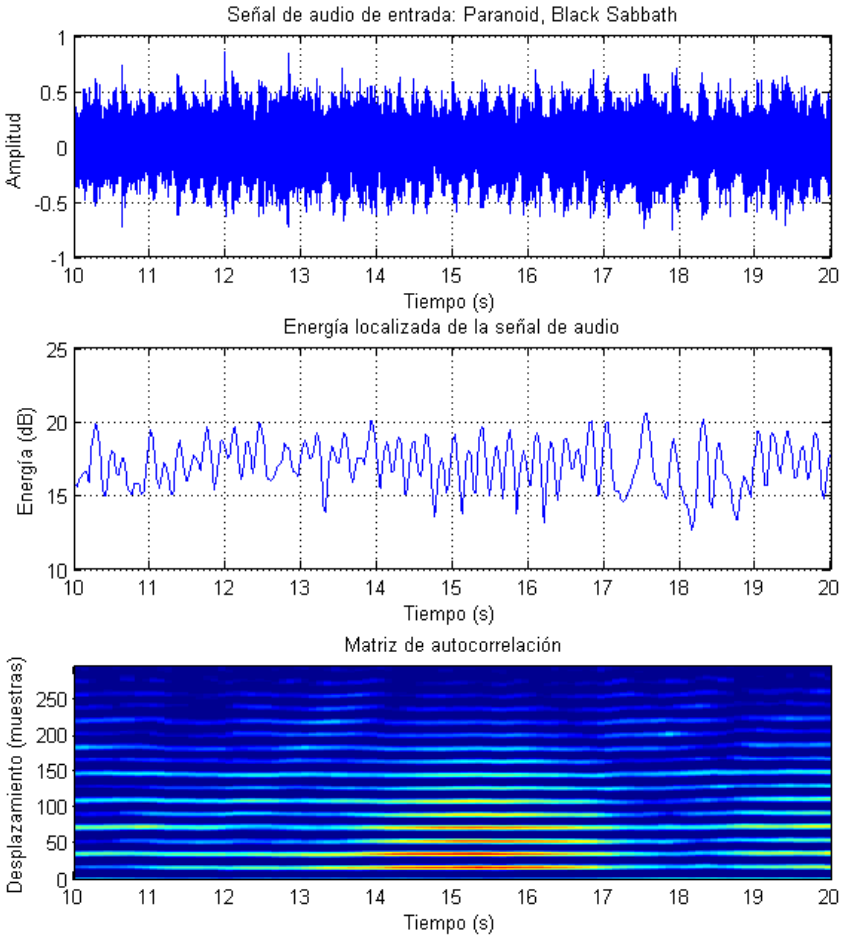

Figura 2-14: Forma de onda, gráfica de energía localizada a lo largo del tiempo y matriz de autocorrelación para la canción "Paranoid" de Black Sabbath, extraída de [12].

La presencia de un ritmo se puede observar en la gráfica de la energía, donde se observa un pulso rítmico marcado. Por otro lado, otra consecuencia se puede observar en la matriz de autocorrelación. Las franjas horizontales que se pueden observar en la matriz son los máximos locales de estas funciones de autocorrelación. Dado que estos máximos locales tienden a mantener su posición a lo largo del tiempo, se puede decir que existe un ritmo claro y de tempo constante.

Por contraposición, se puede observar ahora un segmento de audio sin música -con voz hablada en este caso-, para observar sus posibles diferencias.

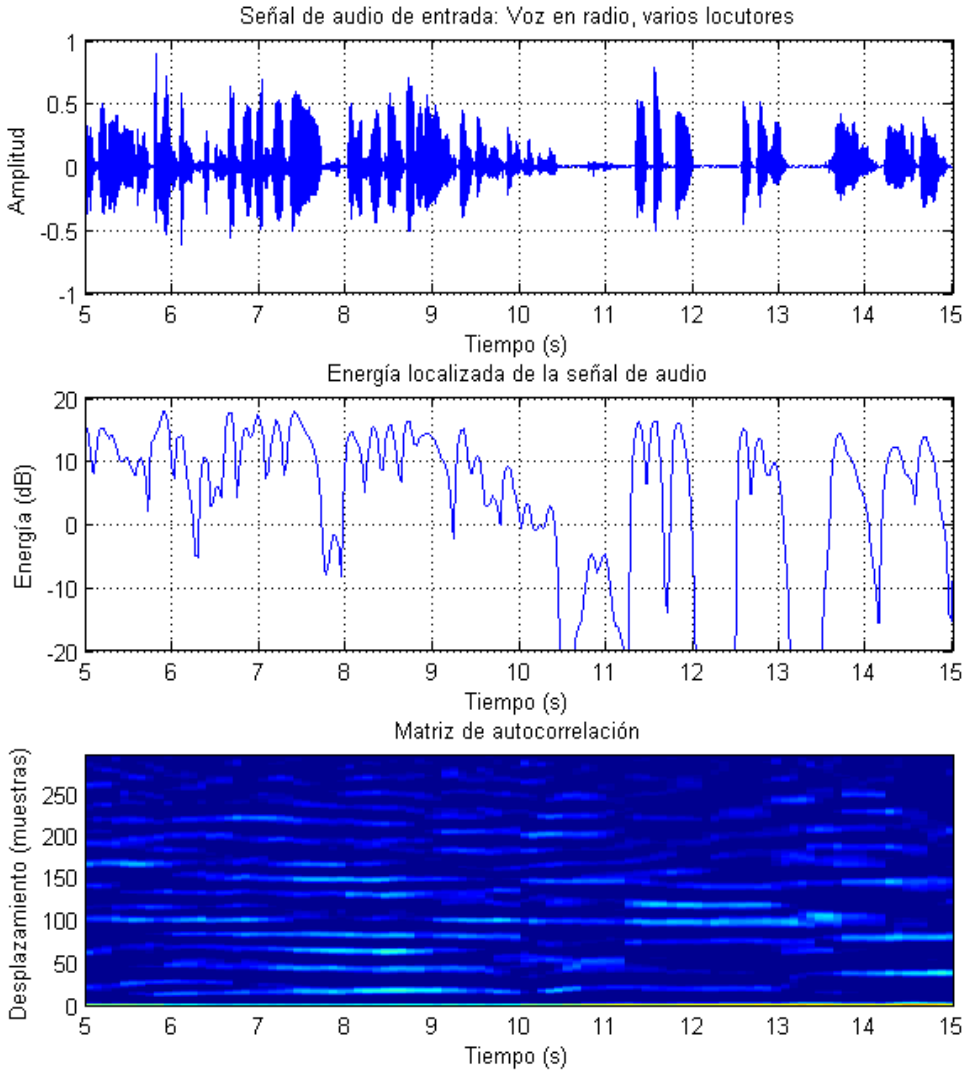

Figura 2-15: Forma de onda, su energía localizada a lo largo del tiempo y matriz de autocorrelación para un segmento de voz, extraída de [12].

En este caso, lo que se observa es que, aunque no hay un ritmo marcado en la matriz de autocorrelación, sí que hay pequeños intervalos de tiempo en los que la energía tiende a seguir patrones periódicos, lo que se traduce en franjas horizontales que se podrían encontrar en un segmento musical. La diferencia fundamental es que éstas aparecen con menos regularidad e intensidad.

2.1.9.2 Detector de armonía

En cuanto al detector de armonía, su objetivo es el de diferenciar las componentes frecuenciales existentes en segmentos de música vs los existentes en los segmentos de voz. A modo de ejemplo, se puede observar la siguiente figura con dos espectrogramas, una canción y un segmento de voz hablada con varios locutores:

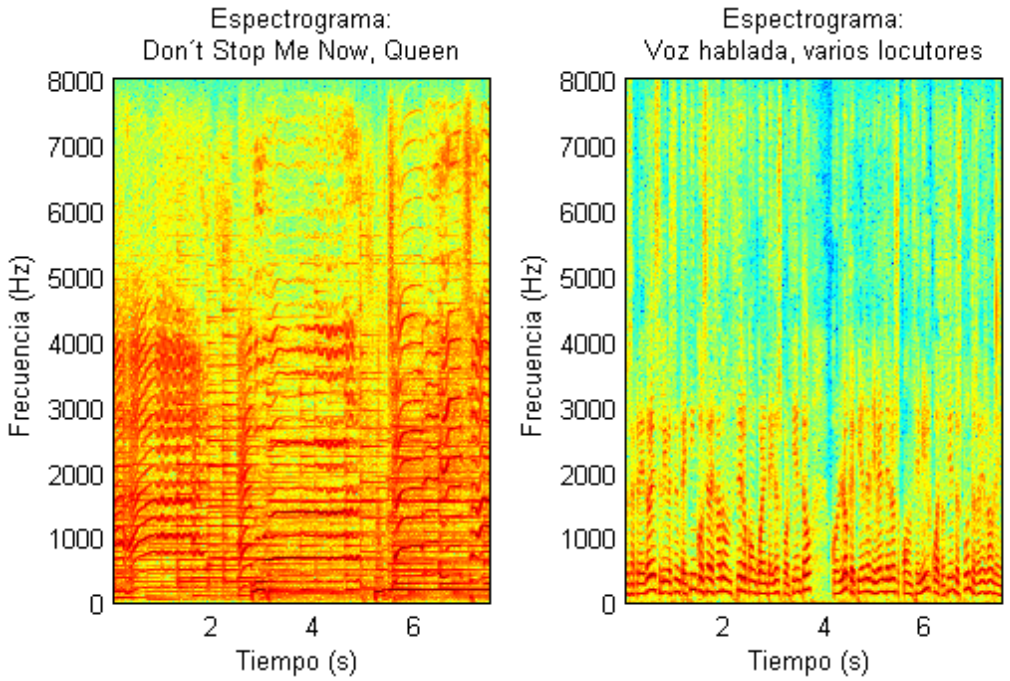

Figura 2-16: Espectrogramas de la canción "Don't Stop Me Now" de Queen (izqda.) y de una conversación con varios locutores (dcha.), extraída de [12].

Esta información permite obtener los llamados cromagramas, que nos permiten resumir estos espectrogramas en doce valores -correspondiéndose cada uno con una nota musical-. La presencia de armonía se vuelve muy evidente en la representación de los cromagramas, tal y como se puede observar en la siguiente figura:

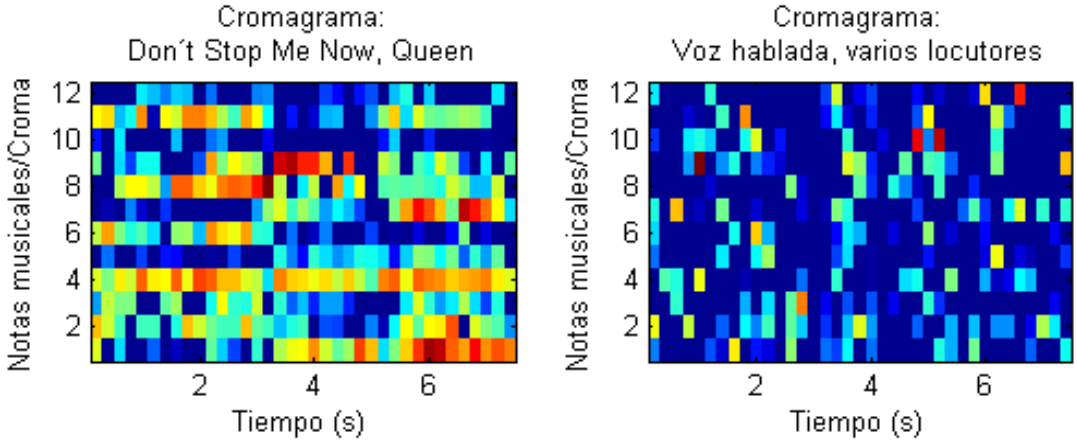

Figura 2-17: Cromagramas de la canción "Don't Stop Me Now" de Queen (izqda.) y de una conversación con varios locutores (dcha.), extraída de [12].

Una vez visto estos valores, [12] fija un umbral para la distinción entre música y no música. Dicho umbral se calcula como:

$$Umbral = min(col) + 0.8 * [\max(col) - \min(col)]$$

Siendo col el vector que contiene los valores en la columna del cromagrama.

2.1.9.3 Fusión de ambos detectores

Con los resultados de ambos detectores ya calculados se ha de tomar una decisión conjunta en base al resultado de los mismos. La razón subyacente para la combinación de los resultados de ambos radica en que se pueden dar señales musicales con poca intensidad rítmica o música en la que la armonía no nos arroja un resultado concluyente.

La herramienta propuesta en [12] para la combinación de ambos resultados es la gráfica de dispersión. Ya que ambos detectores devuelven una señal en función del tiempo con su puntuación, se podrá hacer una gráfica de dispersión bidimensional que represente el valor de ambas puntuaciones en el tiempo.

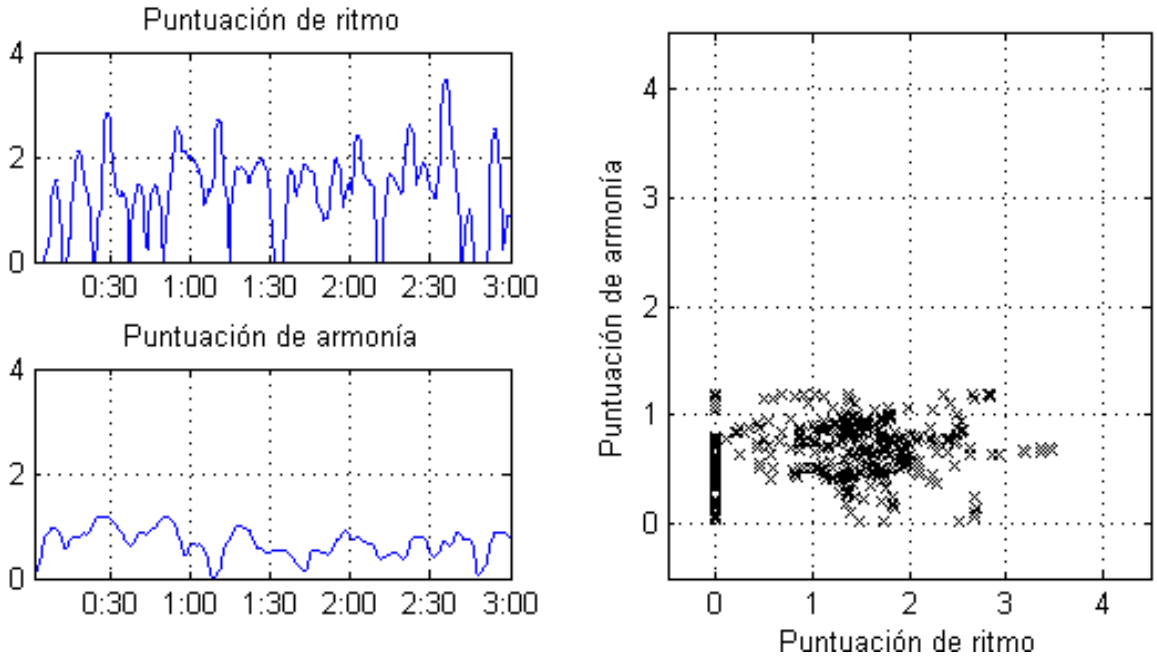

Figura 2-18: Gráfica con la puntuación de ambos detectores del Music Activity Detector (izqda.) y gráfica de dispersión para una señal de voz, extraída de [12].

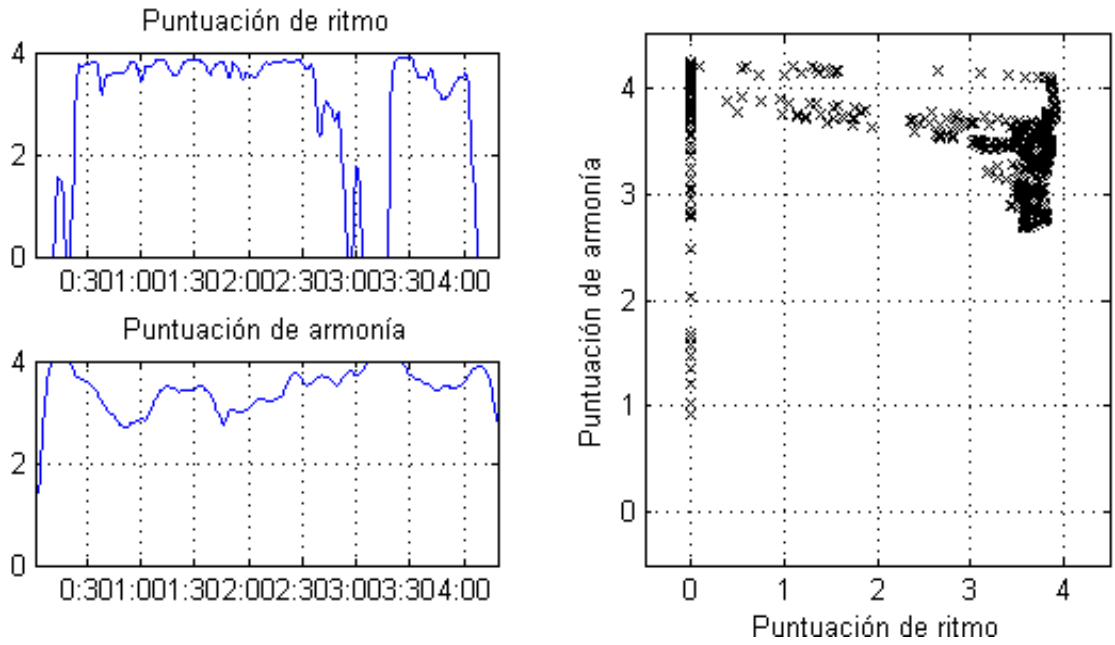

Figura 2-19: Gráfica con la puntuación de ambos detectores del Music Activity Detector (izqda.) y gráfica de dispersión para la canción "Clocks" de Coldplay, extraída de [12].

Finalmente, se calcula una puntuación unidimensional que será la distancia euclídea de cada punto al origen de la gráfica de dispersión.

$$d = \sqrt{\alpha_c * Punt_c^2[n] + \alpha_r * Punt_r^2[n]}$$

Donde $Punt_c^2[n]$ y $Punt_r^2[n]$ son las puntuaciones obtenidas para la muestra n y los valores de alpha son factores de ajuste.

Fijando un valor de distancia de 1,9 como umbral de decisión, este sistema obtiene un total de aciertos de cerca del 97% [12].

2.1.10 Fusión SAD+MAD

Una vez obtenidas decisiones de ambos sistemas se obtiene la siguiente fusión de sistemas:

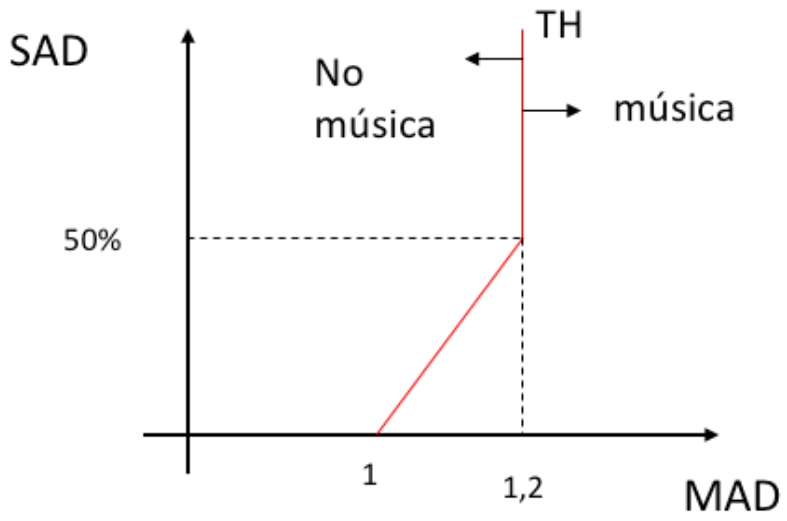

Figura 2-20: Gráfica de decisión del MAD influenciada por SAD.

Donde podemos ver que la puntuación del SAD se tiene en cuenta en la decisión del MAD. De tal forma que, cuando la puntuación del SAD es menor al 50%, es más probable que nos enfrentemos a un segmento de música, por lo que bajamos el umbral del sistema MAD con una cierta pendiente.

Si la puntuación del SAD es muy alta, entonces la puntuación del MAD tiene que estar por encima de 1,2, en lugar de 1, para que el segmento sea considerado música.

2.2 Deep Neural Network

El otro sistema que estudiaremos como alternativa a la integración de SAD + MAD es uno basado en una *deep neural network*, o redes neuronales profundas.

Las redes neuronales nacieron como un intento de replicar sistemas nerviosos biológicos, estos están formados por una serie de nodos, llamadas neuronas, interconectadas unas con otras.

Según [6], una posible clasificación para los modelos de redes neuronales podría ser:

- Modelos inspirados en biología.
- Modelos artificiales aplicados.

En cuanto al primer tipo, las neuronas recibirían señales de otras por medio de sus conexiones y, en función de la señal recibida, enviarían otra señal a otra neurona. En el caso del sistema nervioso humano, se estima que consta de más de cien mil millones de neuronas y que cada una tiene unas mil conexiones de entrada y salida para la comunicación con otras neuronas.

De esta manera, quedaría ese nodo o neurona, como elemento base para la computación del sistema. Quedando algo parecido a:

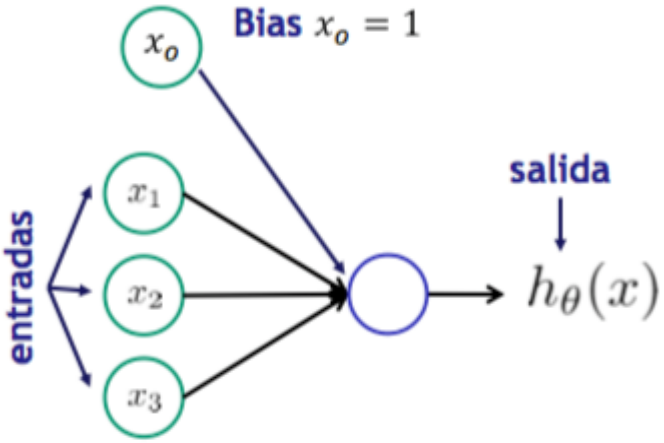

Figura 2-21: Red neuronal de una sola neurona, extraída de [6].

Es importante tener en cuenta que cada sinapsis -conexión de cada neurona- con el resto de la red tiene un peso -o ponderación- que será ajustado a la hora de entrenar la red. Este peso es el llamado vector θ que, en este caso, tendrá dimensión 4 -un valor por cada conexión-. La salida dependerá de la función de activación que tenga la neurona en sí, en nuestro caso, se utiliza en todas ellas la función sigmoide que tiene la forma:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Añadir capas intermedias -capas ocultas- a la red neuronal es habitual en la bibliografía, y más para el campo de VAD, el resultado quedaría similar a:

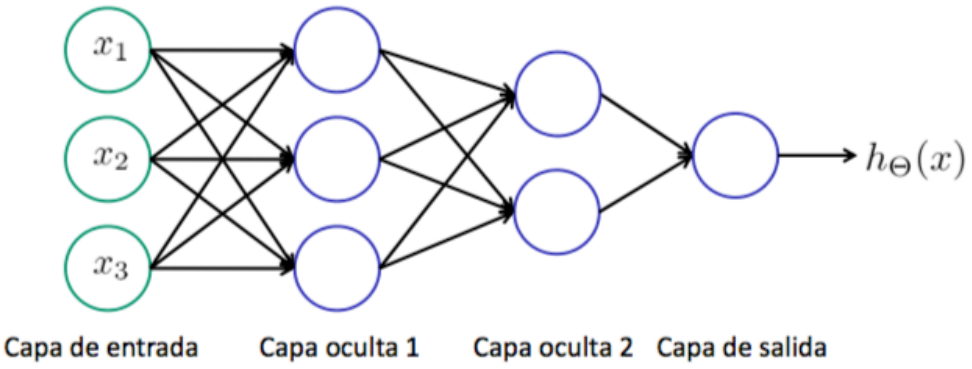

Figura 2-22: Red neuronal formada por dos capas ocultas, extraída de [6].

Además, en esta figura se puede observar un hecho no muy habitual en las redes neuronales pero que, sin embargo, sí que presenta la red neuronal que usaremos en este proyecto.

Si observamos la capa oculta 2, veremos que su número de neuronas es menor al número de neuronas de la capa oculta 1, este hecho se denomina como "*Bottleneck* en la capa 2". La implicación que tiene este hecho es que, antes de decidir la salida, ya se está haciendo un filtrado de las características que serán relevantes a la hora de la clasificación final que tendrá que hacer la red neuronal.

La DNN estudiada a lo largo de este proyecto ha sido la misma que la que el grupo Audias-ATVS ha presentado a NIST OpenSAT, disponible en [3]. Este sistema está basado a su vez en [4].

Las DNN han demostrado poder dar buenos resultados en la tarea de *Voice Activity Detection*, tal y como podemos ver en [5]. Sin embargo, en este caso, no nos vamos a centrar en la solución desarrollada en [4], que utiliza una DNN con una memoria presente en cada neurona que la conforma -llamada CLDNN-.

La DNN utilizada en este sistema consta de tres capas ocultas de quinientas neuronas cada una, con una función tipo sigmoide como función de activación. La capa de salida consta de dos neuronas, una por clase -voz y no voz-. En cuanto a la entrada, cada *frame* es analizado con los 15 anteriores y 15 siguientes *frames*.

Además, el sistema toma para cada *frame* los 20 primeros MFCCs. Todo esto se traduce en que el vector de entrada a la DNN tendrá dimensionalidad 620.

El entrenamiento, por otro lado, se realiza utilizando el algoritmo de descenso por gradiente en *mini-batches* de 512 características, utilizando la entropía cruzada como función de coste del algoritmo.

Por último, cabe destacar que, en el caso de la DNN se realiza un proceso de filtrado de los resultados posterior al resultado de la propia DNN. Esto significa que, una vez obtenido el resultado de la DNN para la totalidad de los segmentos de audio, se mirará también los segmentos cercanos en el tiempo, con el fin de eliminar efectos que, puede que en puridad estén ahí -por el proceso de coarticulación del habla, por ejemplo-, pero que en el etiquetado no estarán. Esto es, son silencios de muy corta duración presentes de manera natural en el habla, pero que no son tenidos en cuenta en el etiquetado.

En el apartado de pruebas se probará con distintos valores de este filtrado, con el objetivo de ver cuál es el óptimo entre ellos.

2.3 Evaluación NIST OpenSAT

OpenSAT es una evaluación de organizada por NIST -*National Institute for Standards and Technology*-. La evaluación de la primavera de 2017 constará de tres tareas: *Speech Activity Detection, Keyword Search y Automatic Speech Recognition.*

El objetivo de esta evaluación es fijar un criterio base de rendimiento para las herramientas existentes de análisis de voz, cuándo éstas son sometidas a muchos datos complejos.

En cuanto a los datos de entrenamiento, la evaluación deja utilizar cualesquiera datos que estén disponibles de manera pública para desarrollar y entrenar los sistemas. Aunque se ha de tener en cuenta que los datos de entrenamiento deberán ser descritos en la documentación del sistema.

Los datos de desarrollo serán distribuidos a los participantes y podrán ser usados para cualquier tarea, incluyendo desarrollo y entrenamiento del sistema.

De las tres tareas que se evalúan en OpenSAT nosotros nos centraremos en el SAD. El objetivo del SAD, como ya se ha expuesto, es la detección de segmentos donde haya voz en un fichero de audio.

2.3.1 Medidas de Rendimiento

El rendimiento en los sistemas VAD es evaluado mediante dos valores fundamentales: *false alarm* -FA- y *missed speech* -MS-. FA se refiere a momentos en los que el sistema VAD detecta voz, pero, sin embargo, la realidad es que no la hay. Por contraposición, MS se referirá a aquellos momentos en los que el sistema no detecte voz cuando sí que la hay.

Sin embargo, dependiendo de la aplicación concreta ambos errores no tienen por qué ser tenidos en cuenta con la misma importancia. Por ejemplo, para un codificador de voz que se use en telefonía puede no ser grave enviar tramas que no contenían voz -incurriendo así en FA-, pero sí que es grave que no se envíen tramas que sí que contenían voz -incurriendo en MS-.

En el caso de la evaluación NIST OpenSAT, el rendimiento será evaluado de acuerdo con una DCF -*Detection Cost Function*-, este valor dependerá de los valores de *false alarm* y *missed speech* que arroje el sistema, dando más importancia a los valores de MS. El objetivo es, por tanto, ajustar el sistema de forma que se minimice esta DCF.

Dicha DCF se calculará de acuerdo con:

$$DCF = 0.75 \times P_{MISS} + 0.25 \times P_{FA}$$

Donde P_{MISS} es la probabilidad de un falso negativo y P_{FA} es la probabilidad de. Un falso positivo. Están definidas de la siguiente forma:

$$P_{MISS} = \frac{Tiempo\ total\ de\ Miss}{Tiempo\ total\ de\ speech}$$

$$P_{FA} = \frac{Tiempo\ total\ de\ Falsa\ Alarma}{Tiempo\ total\ de\ speech}$$

Conviene recalcar otra vez que, tal y como se puede observar, el cálculo de DCF penaliza mucho más gravemente tener un falso negativo que un falso positivo. Esta circunstancia se tendrá que tener muy en cuenta a la hora calibrar nuestro sistema.

Por otra parte, la evaluación tiene en cuenta errores en el propio etiquetado realizado por humanos, esto es, la resolución del etiquetado es finita. Para ello se define un "*collar*" un periodo de medio segundo delante y detrás del segmento de voz etiquetado en el que no se calcula el rendimiento del sistema -o no se tiene en cuenta-. Este collar también compensa ambigüedades en las anotaciones de ruidos.

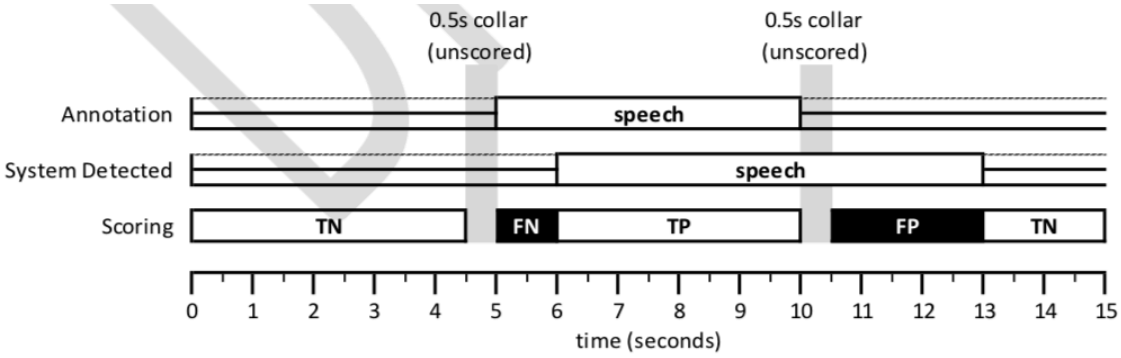

Figura 2-23: Gráfico explicativo de las medidas de rendimiento de NIST, extraída de [11].

3 Entorno experimental

3.1 Etiquetado

Como primera etapa de este trabajo se decidió hacer un etiquetado propio de la base de datos con la que se trabajaría en la evaluación de NIST OpenSAT. Dicha base de datos está compuesta por tres subconjuntos:

- Babel, compuesta por conversaciones telefónicas con ruido.
- SSSF, audios provenientes de conversaciones entre servicios de emergencias.
- VAST, audios procedentes de redes sociales.

3.1.1 Estudio previo

Con el objetivo de decidir todo el detalle del etiquetado que se realizará un primer estudio del contenido de los distintos conjuntos dentro de la base de datos. En general, se busca poder saber con más detalle su composición para hacer un etiquetado lo más útil posible tanto para esta como futuras aplicaciones.

3.1.1.1 Babel

Tal y como ya se ha apuntado Babel, está compuesta por conversaciones telefónicas con ruidos de distinta naturaleza. Estas conversaciones están habladas en idioma pastún.

La forma en la que los audios son entregados es separados en dos canales, esto es, de los terminales que participan en la conversación, cada uno tiene su propio fichero de audio. Esto hace que, por ejemplo, momentos de silencio de un canal tengan presencia de conversación por parte del otro canal.

Es por esto, que el primer paso consistió en la realización de un script de Matlab que uniese ambos canales de la misma conversación en uno solo. En nuestra aproximación, concretamente se colocó el canal catalogado como *"inline"* en el canal izquierdo de un audio estéreo y, por el contrario, el canal *"outline"* en el canal derecho.

Conviene destacar también, que no todos los *inline* tienen su correspondiente *outline*, por lo que, para esa clase de archivos, no tendremos la información completa de la conversación.

Las características observadas han sido:

- Audio de muy mala calidad, incluso teniendo en cuenta que se trata de voz telefónica. Esta circunstancia se ha observado en, alrededor de, el 50% de las grabaciones.
- Pequeñas interrupciones entre locutores. Circunstancia muy habitual.
- Silencios. Bastante habitual y, como es de esperar, con una presencia muy importante en aquellos archivos que no tienen el canal *outline* presente.
- Risas sin carcajadas, presencia esporádica.

- Voces muy bajas. Por ejemplo, alguien deja el teléfono -o lo aparta de la cara-, para preguntar a alguien que está próximo a él, para luego retomar la conversación telefónica.
- Por el resto, mayoritariamente se trataban de conversaciones telefónicas estándar.

Esta parte de la base de datos está dividida en parte de desarrollo -*dev*- y en parte de evaluación -*eval*-. La cantidad de audio disponible se puede observar en la siguiente Tabla:

	Número de ficheros	**Duración total**
DEV	15.637	214:38:53,63
EVAL	198	30:38:46,60

Tabla 3-1: Duración de la Base de Datos BABEL.

3.1.1.2 SSSF

Estos audios se corresponden con conversaciones entre servicios de emergencias. El idioma de las conversaciones es inglés en todos los casos. Esta parte de la base de datos se caracteriza por tener multitud de segmentos con un nivel de ruido que hace casi inaudible lo que se está diciendo -por ejemplo, un bombero utilizando una sierra-, combinado con partes claramente inteligibles -por ejemplo, conversación con el centro de control-.

Las características observadas han sido:

- Voz de radio policial. Con una presencia mayoritaria en el audio.
- Pitidos presentes en el ambiente. Con una presencia muy esporádica.
- Silencios muy largos. Algo habitual.
- Ruidos/voces extremadamente ruidosas combinando con segmentos limpios. Esta circunstancia es bastante habitual.
- Presencia de muchos pitidos. Probablemente pitidos de sincronismo para los sistemas utilizados.
- Voces con ellas mismas retrasadas. Por producirse en un lugar con mucho eco u oírse un sistema de megafonía. Esta circunstancia es poco habitual.
- Se puede oír en algún momento sirenas.
- Es habitual encontrar segmentos de varios segundos sin locución, pero con mucho ruido de fondo.
- Voz con distorsión muy alta.

	Número de ficheros	**Duración total**
DEV	6	0:30:02,00
EVAL	6	0:31:55,00

Tabla 3-2: Duración de la Base de Datos SSSF.

3.1.1.3 VAST

En cuanto a VAST, esta parte de la base de datos está formada por audios provenientes de redes sociales de Internet, principalmente, *Youtube*. En este caso, los idiomas presentes en los audios son:

- Inglés.
- Mandarín.
- Árabe.

Es importante tener en cuenta que, dada la procedencia de estos audios, las situaciones que nos encontramos son tremendamente diversas. No se pueden establecer unas líneas generales en cuanto al audio porque, simplemente, no las hay. Cada audio es tremendamente distinto del anterior.

Sin embargo, en este estudio previo se pudieron observar las siguientes circunstancias:

- Voz con muchos ecos, situación bastante habitual.
- Conversaciones normales entre dos o más personas. También bastante habitual.
- Palmadas o similar, esto es, ruidos muy efímeros, pero con energía alta.
- Voz telefónica. Casi inexistente.
- Ladridos de perros o el llanto de un perro. Casi inexistente.
- Silencios largos, habitual.
- Voces lejanas o hablando detrás de la cámara. Habitual.
- Conversaciones en entorno ruidoso, como una cafetería o similar. Habitual.
- Locución muy rápida y limpia con pequeños ruidos muy impulsivos. Carácter muy esporádico.
- Voces de niños pequeños con comentario más cerca de cámara. Da la sensación de que un adulto está grabando y, a la vez, está manteniendo una conversación con el niño, de tal manera que, hay una voz en primer plano -la del niño-, y otra tras la cámara -la del adulto-. Esta circunstancia era muy habitual.
- Voces de fondo, como en un parque de atracciones o similar. Esta situación era esporádica.
- Gritos, esporádico.
- Audios de vídeos en entorno controlado -con voz muy limpia y cercana a micrófono-. Habitual.
- Ruidos continuos, tales como viento, un motor, o claxon sonando durante toda una grabación. El caso del viento se da en un número limitado de archivos, pero aquel audio en el que aparece está en la totalidad del mismo. El resto de ruidos es esporádico.
- Audios de introducciones, esto es, voz en off más música. Esporádico.
- Música grabada directamente. Esporádico.
- Cantando con piano. Esporádico.
- Ruidos de animales en primer plano como algo habitual.
- Toses, bastante esporádico.
- Ruido de agua -ruido tipo grifo de agua abierto-, muy esporádico.
- Voz principal que destaca muy poco sobre voces de fondo -esto es, grabaciones de gente lejana a quienes apenas se oye sobre el fondo-. Es habitual esta circunstancia.

- Conversaciones con conversaciones de fondo como ruido e, incluso risas de fondo. Esporádico.

Como se puede observar, se dan muchísimas circunstancias, algunas más habituales que otras, pero es imposible acotar mucho el análisis, debido a la poca limitación por parte de las redes sociales de tipo de contenido y/o situación de grabación.

	Número de ficheros	**Duración total**
DEV	300	13:50:52,99
EVAL	200	9:14:40,09

Tabla 3-3: Duración de Base de Datos VAST

3.1.2 Conjunto final por etiquetar

Una vez estudiados todos los subconjuntos de la Base de Datos se debe elegir un subconjunto para nuestro etiquetado. Para la decisión se siguieron los siguientes razonamientos:

- Conjunto de audio lo más diverso posible.
- Máximo número de archivos posible.
- Duración aproximada del audio de 4 horas.

Estos requisitos se tomaron por lo siguiente: tener un conjunto que represente el máximo número de situaciones posibles y que la duración de la tarea del etiquetado no fuese excesiva.

Por ello, se llegó a la conclusión de que se elegirían los ficheros más cortos del conjunto de desarrollo -dev- de VAST, hasta completar las cuatro horas. El resultado fue:

	Número de ficheros	**Duración total**
DEV	123	3:58:13,21

Tabla 3-4: Duración total de audio para etiquetar.

3.1.3 Niveles de etiquetado

La propuesta que se hace de etiquetado consta de tres niveles:

1. Ruido ambiente.
2. Ruidos esporádicos.
3. Voz.

Para el nivel de ruido ambiente se colocar las siguientes etiquetas:

- Voces (cafetería)
- Música HQ
- Música ambiente
- TV Fondo o Radiodifusión
- Fondo genérico
- Silencio (como NO)

En cuanto al nivel de etiquetado de ruidos esporádicos estarán las siguientes etiquetas:

- Risas
- Pitidos/tonos
- Golpe
- Animales
- Palmas
- Claxon

En general, la distinción entre un ruido ambiente y uno esporádico será la duración del mismo. Poniendo de criterio que una duración de más de cinco segundos es ambiente y menos, esporádico. Sin embargo, también se hace distinción en nivel de energía, por ejemplo, una motosierra que cope durante 20 segundos la totalidad de un audio, no se considerará ambiente si no esporádico.

Relativo al nivel de etiquetado de voz, este se trata de un nivel mucho más completo que los anteriores. Dentro del propio nivel habrá tres subniveles por debajo, que serán: calidad, plano de locución y hablante.

	Calidad	Plano	Hablante
VOZ	HQ	1 / 2	a,b,c,d… Reservada la "n" para "no identificado"
	Telefónica	1 / 2	
	Radio -Walkie-	1 / 2	
	Alta distorsión	1 / 2	

Tabla 3-5: Detalle del nivel de etiquetado VOZ

De esta manera, una locución en alta calidad, presente en primer plano del hablante "c", será etiquetada como:

HQ1C

3.1.4 Herramienta de etiquetado

La herramienta que utilizaremos será *Wavesurfer*. Para su correcta utilización será necesaria una configuración previa de la misma para definir los niveles de etiquetado y las etiquetas finales que conformarán cada nivel.

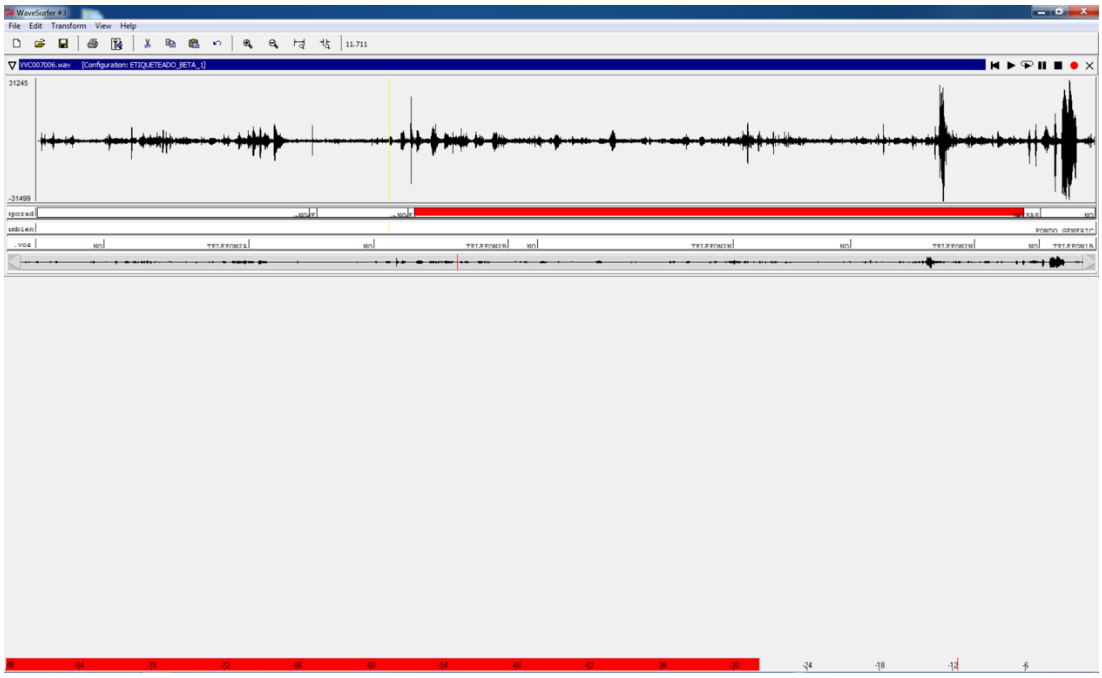

Figura 3-1: Herramienta para el etiquetado Wavesurfer

Para la configuración de la herramienta se deberá hacer *click* derecho, *click* en *"Create Pane"* y elegir tipo *"Transcription"*. Una vez tenemos creado los tres niveles que queremos, tendremos que ajustar sus propiedades para tener tanto las etiquetas predefinidas que nos interesan como el formato correcto. Nosotros hemos elegido el formato HTK.

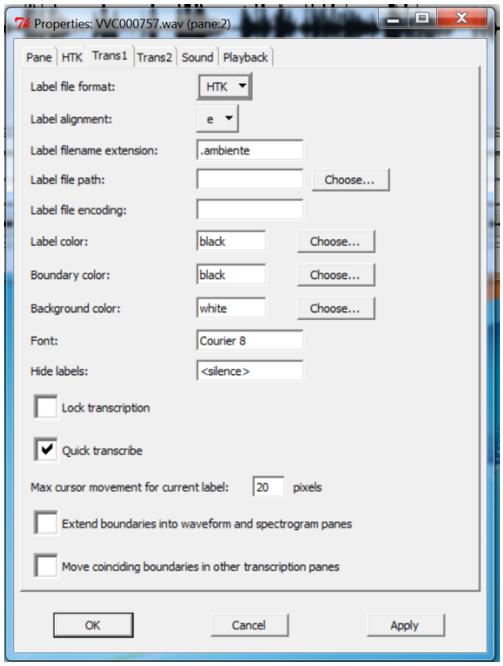

Figura 3-2:Configuración de las propiedades Wavesurfer

Una vez completada la configuración, se podrá guardar la misma y podrá ser utilizada, una vez copiada a la carpeta de configuración de *Wavesurfer*:

C:/Users/Usuario/.wavesurfer/1.8/configurations

El formato que hemos utilizado, tal y como se puede ver en la Figura 3-2, es el HTK. La razón de la utilización de este formato en concreto es que es el mismo que se usó en [1], [7], [8] y [9] para la realización de la Base de Datos "Audias-ATVS Radio".

El formato HTK es un formato muy utilizado actualmente como formato de etiquetado de audio. Consta de tres filas y tantas columnas como etiquetas haya presentes en el archivo. Un ejemplo de fila sería:

223904739 309845930 HQ2C

Esta fila se correspondería con una etiqueta empieza en el momento que marca el primer campo, finaliza en el momento que marca el segundo campo y es del tipo que marca el tercer campo. Hay que tener en cuenta que la magnitud de los momentos tanto inicial como final están con 7 decimales de segundo.

En suma, el anterior ejemplo se correspondería con una locución en alta calidad, segundo plano del locutor C, que empieza en el segundo 22,3904739 y acaba en el segundo 30,9845930.

El resultado total de esta configuración será de tres archivos de etiquetas por cada archivo .wav. Estos archivos de etiquetas estarán en formato HTK y tendrán las extensiones:

- ".ambiente" para las etiquetas del ruido ambiente.
- ".esporadico" para las etiquetas de los ruidos esporádicos.
- ".voz" para las etiquetas del nivel de voz.

3.2 Datos OpenSAT

Los datos arrojados por cada uno de los sistemas han sido obtenidos directamente de los archivos disponibles del grupo Audias-ATVS para la evaluación NIST OpenSAT. Estos datos se componen de un fichero .tsv por cada archivo de audio de la base de datos en el que se da las decisiones del sistema sobre si un segmento es "voz" o "no voz". Un ejemplo de una línea de un archivo podría ser:

OpenSAT17_VAST_Dev OpenSAT17 OpenSAT17_VAST_Dev SAD VVC000197 8.25 20.74 speech 0.72

Donde:

- Los cuatro primeros campos son datos del experimento.
- El quinto campo -VVC000197- es el archivo.
- El sexto campo es el momento inicial en segundos.
- El séptimo campo es el momento final en segundos.
- El octavo campo es la decisión entre *"speech"* y *"no speech"*.
- El noveno campo es la confianza en la decisión.

Estos datos están disponibles para:

- El sistema SAD.
- El sistema SAD+MAD.
- El sistema basado en DNN.
- El sistema basado en la fusión presentada en NIST OpenSAT de SAD+MAD y la DNN.

3.3 Entrenamiento de la DNN

Debido a la selección realizada para el etiquetado en este trabajo, hubo que reentrenar la DNN ya que el subconjunto de audio que se podía utilizar como medida de rendimiento había sido ya utilizado para su entrenamiento.

El primer paso para el entrenamiento de la DNN consiste en realizar la lista de archivos que serán utilizados para dicha tarea. En nuestro caso carece de sentido entrenar la DNN con datos de SSSF sabiendo sus características tan peculiares y, a la vez, tan distintas, con respecto a nuestro subconjunto de VAST.

Por esta razón la lista de archivos para el entrenamiento de la DNN se corresponde con el 85% del conjunto de datos de BABEL conversacional y el 85% de las horas de audio de VAST que no serán utilizadas como elemento de test.

El 15% restante de horas de VAST se dedicará como conjunto de validación de la DNN. El conjunto de validación, dentro de un entrenamiento para una DNN, consiste en un subconjunto de datos, separados de los de entrenamiento, que sirven para calibrar, desde un punto de vista sin sesgos, la bondad del entrenamiento.

El total de archivos utilizados para el entrenamiento de la DNN, entre BABEL conversacional y VAST es de 1178. Cabe destacar que, a la hora de entrenar la red, es importante no entregar los archivos ordenados, es decir, no tiene sentido entregar primero todo el conjunto de BABEL y luego VAST o viceversa. Es por esto por lo que, antes de entrenar la red, se hace una aleatorización de la lista de los ficheros a utilizar.

En el caso de validación el total de archivos es de 163, entre BABEL y VAST.

4 Pruebas y resultados

4.1 Estadísticas de la Base de Datos

De cara a realizar una adecuación de los sistemas al corpus bajo estudio, es importante realizar un análisis estadístico sobre las distintas situaciones a las que se enfrentarán los sistemas durante su ejecución.

El primer dato que nos ha interesado es el de hablantes por fichero. La distribución de los mismos a lo largo de los ficheros se muestra en la siguiente figura:

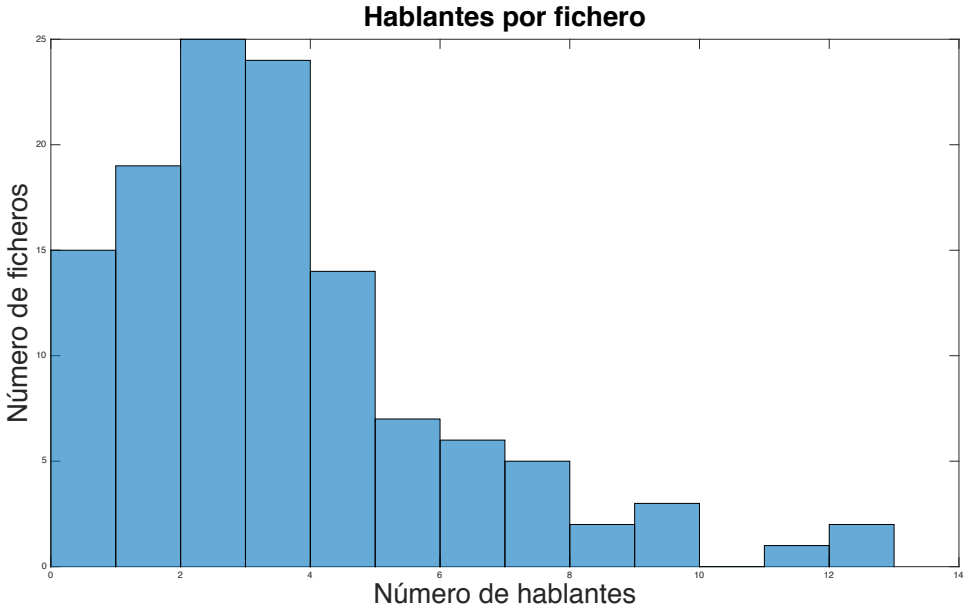

Figura 4-1: Histograma del número de hablantes por fichero.

Calculando la media, se obtiene que, en este corpus, hay 3,0976 hablantes por cada fichero de audio.

También se ha calculado la duración media de las locuciones por cada archivo. La media, en este caso, es de 5,62 segundos. Sin embargo, dada la naturaleza tan diversa de la Base de Datos, es importante ver cómo quedan dichos resultados en un histograma:

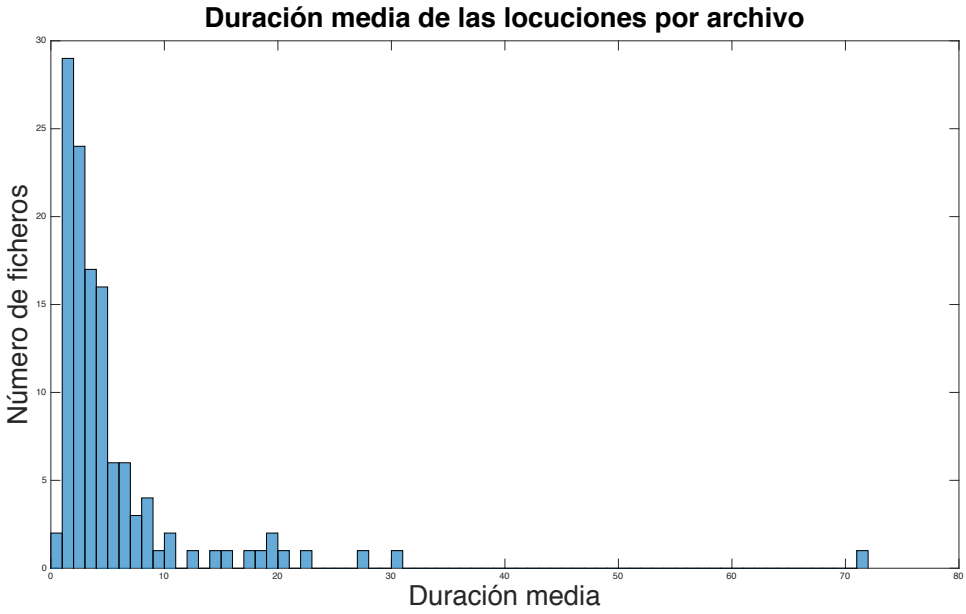

Figura 4-2: Histograma de la duración media de las locuciones por archivo.

Tal y como se puede observar en la Figura 4-2, en general las locuciones tienen duraciones cortas, ya que vemos que, en la mayoría de los casos, no se superan los 10 segundos de locución. Tan solo hay un espurio, que se corresponde con un archivo en el que el audio está muy controlado y, en su totalidad, es un locutor hablando directamente al micrófono.

Una característica que puede tener un alto grado de importancia en un sistema SAD, es el plano de locución. De nuevo, debido a la naturaleza tan diversa de la Base de Datos, los resultados esperados son que habrá una gran cantidad de los archivos que tendrán locución en ambos planos. Los resultados se pueden ver en la siguiente tabla:

Número total de archivos	123
Locuciones solo en primer plano	38
Locuciones solo en segundo plano	0
Locuciones en ambos planos	79
Sin información de plano de locución	6

Tabla 4-1: Planos locuciones en la Base de Datos.

Una vez realizado el estudio previo del corpus -realizado en el punto 3.1.1-, destaca que el ruido de carácter esporádico más abundante se corresponde con la risa. Este ruido esporádico aparece en el 48,7805% de los ficheros. Distribuidos de la siguiente manera:

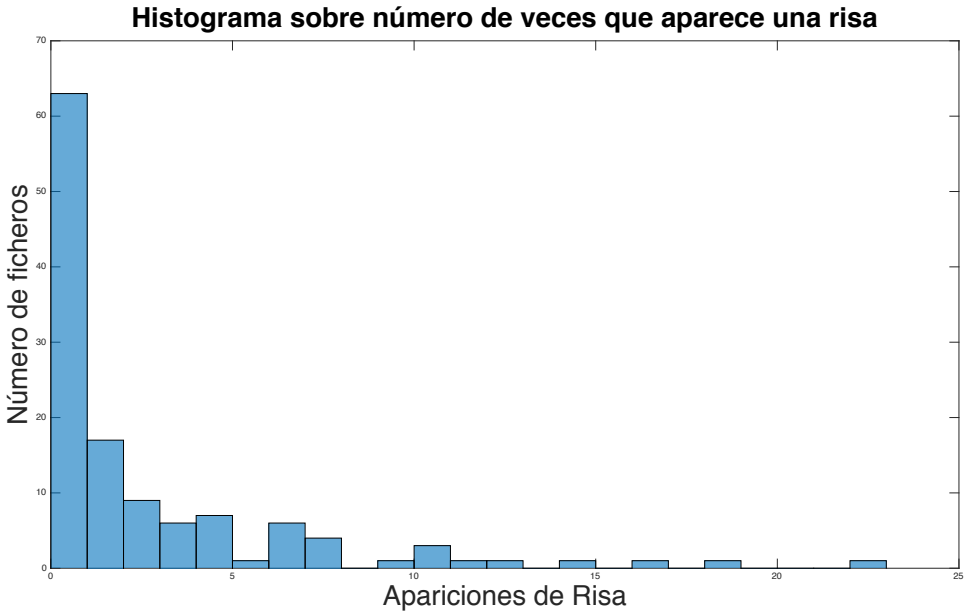

Figura 4-3: Histograma apariciones de risas en los ficheros.

Podemos ver cómo, efectivamente, en 63 ficheros -el 51,2915% de los ficheros del corpus-, no hay risa en ningún momento del audio. También cabe que, generalmente, el número de veces que aparece una risa en un archivo no supera las 10 ocasiones.

Otro aspecto importante de la risa es la duración media de la misma, esto es, influirá de manera distinta una risa de medio segundo que se repita varias veces en el archivo, y una risa que aparece sólo una vez, pero dura varios segundos. También se ha sacado la distribución de la duración media de las risas en cada archivo, quedando:

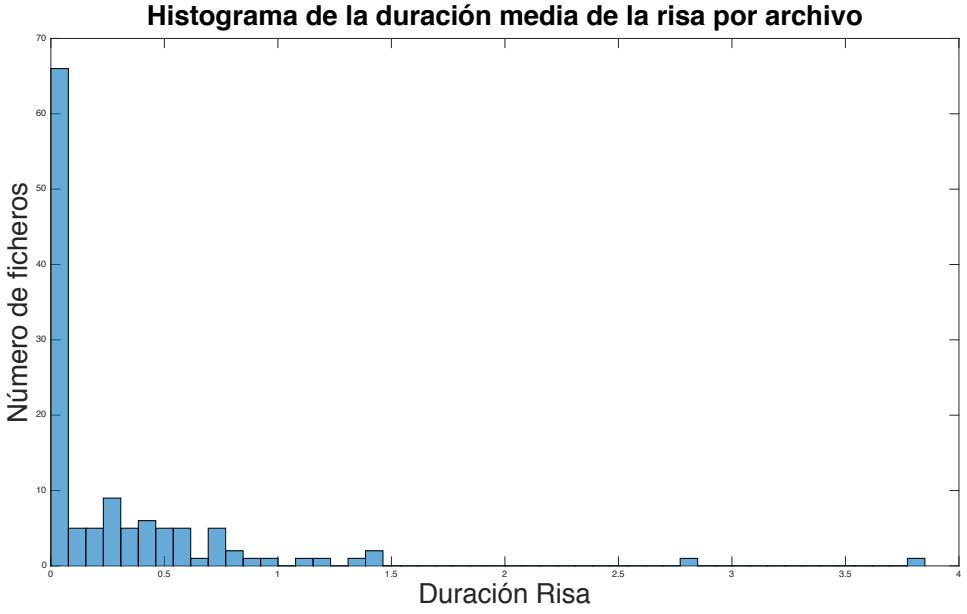

Figura 4-4: Histograma de la duración media de la risa en los ficheros.

Tal y como se puede observar, la duración media de la risa no suele sobrepasar el segundo y medio y, tan solo en un caso, se puede ver una duración media cercana a los 4 segundos.

Cabe destacar también, en el apartado de ruidos esporádicos, que un 30,8943% de los archivos -38 archivos-, no presentaban ruido esporádico de ningún tipo.

En cuanto al nivel de etiquetado de ruidos ambiente, lo más destacado en la observación ha sido la cantidad de solapamiento entre voces de fondo y voces de la grabación, ya sea en primer o en segundo plano. Este nivel de solapamiento se puede ver en los siguientes histogramas:

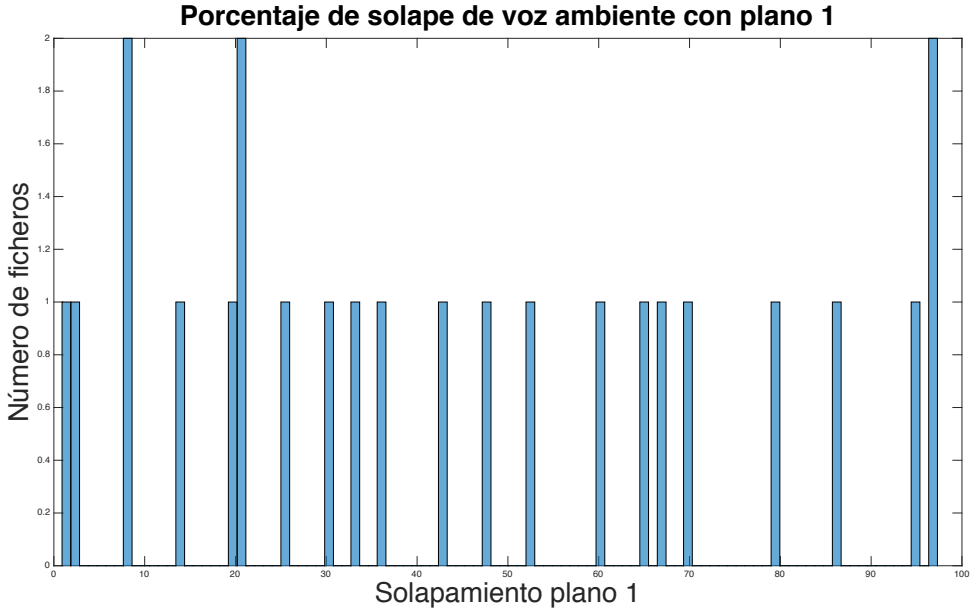

Figura 4-5: Histograma del porcentaje de solape de voz ambiente con plano 1 -por fichero-.

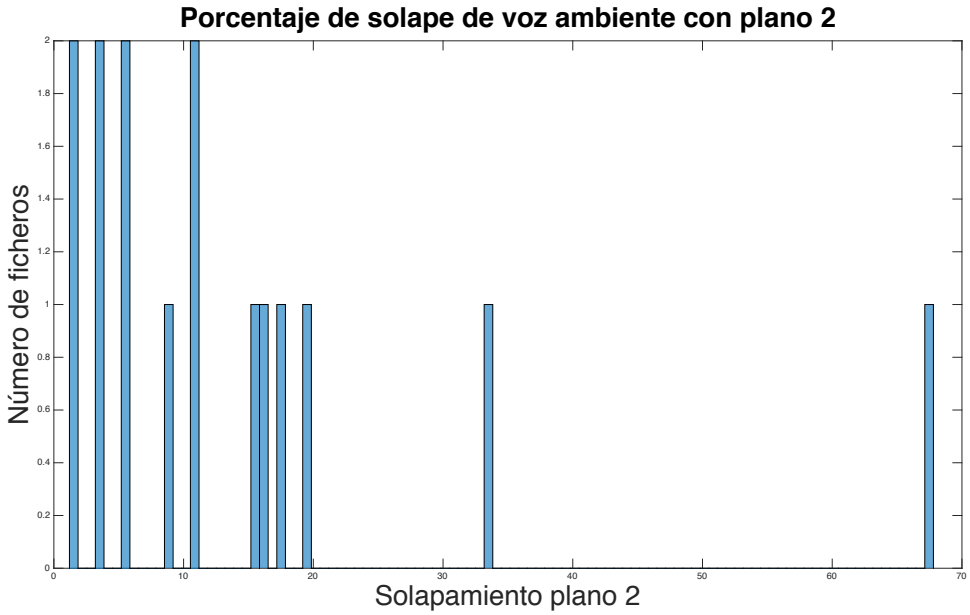

Figura 4-6: Histograma del porcentaje de solape de voz ambiente con plano 2 -por fichero-.

Es importante destacar que, debido a que en la mayoría de los casos no hay solapamiento de ningún tipo, se ha representado en el histograma solo aquellos valores que sí que tienen solapamiento. Esto ha sido así para poder tener un mayor detalle en la representación de los mismos.
En este corpus concreto, los resultados concretos de solapamiento se pueden ver en la siguiente tabla:

Número total de archivos	**123**
Número de archivos con solapamiento en el plano 1	24
Número de archivos con solapamiento en el plano 2	15
Porcentaje de solapamiento plano 1 -temporal-	45,0266%
Porcentaje de solapamiento plano 2 -temporal-	14,8717%

Tabla 4-2: Solapamiento de locuciones con voces de fondo.

4.2 Rendimiento de los sistemas

Con el objetivo de comprobar qué sistema, y en qué circunstancias, arroja mejores resultados, en este punto se estudia el rendimiento de los mismos de acuerdo con los criterios de [11] ya descritos en el punto 2.3.

4.2.1 Rendimiento SAD

En primer lugar, vamos a comparar directamente los porcentajes de MS y FA por archivo, estos resultados se pueden ver en la siguiente figura:

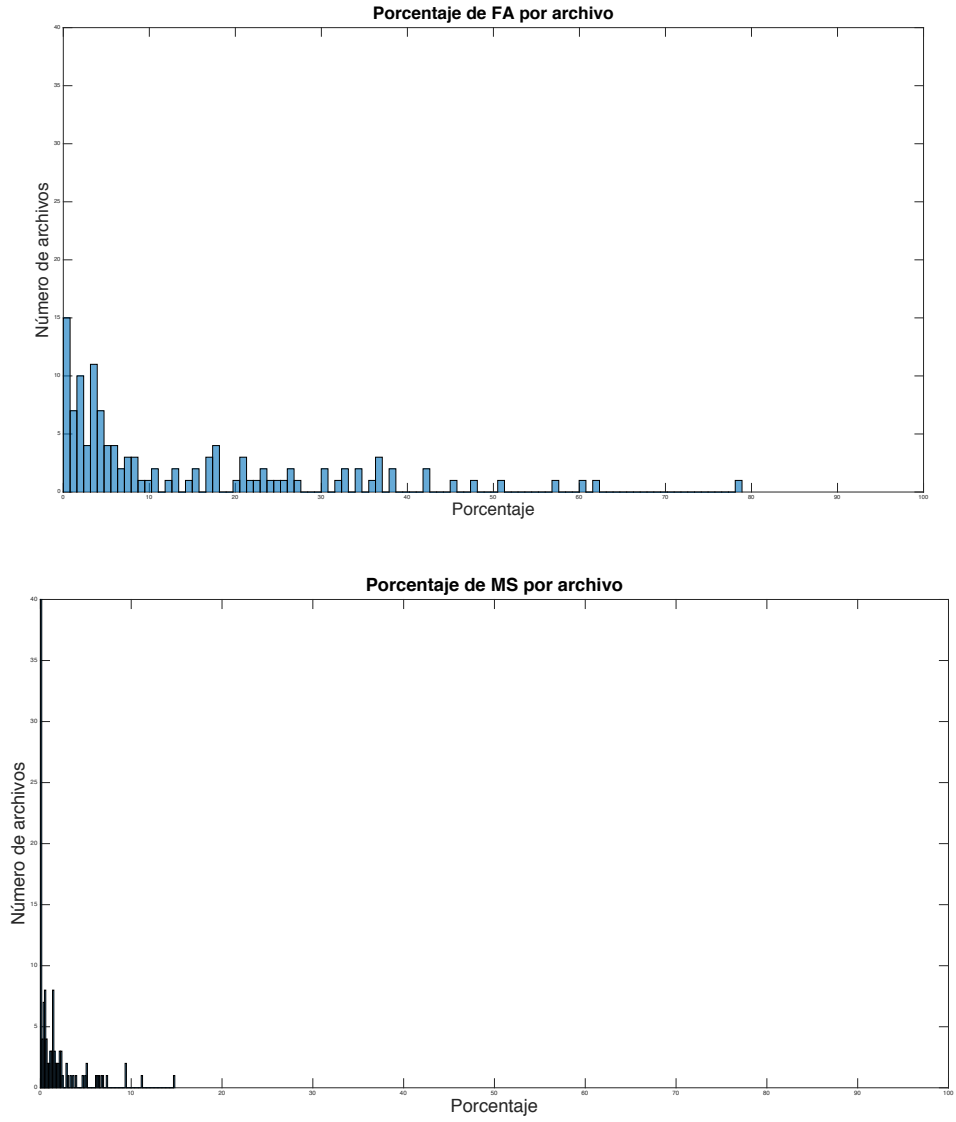

Figura 4-7: Histogramas de False Alarm (arriba) y Missed Speech (abajo) por cada fichero del conjunto etiquetado.

Se puede observar que hay un gran número de archivos -36 archivos- cuyo valor de MS es, directamente cero. Además, el error medio de FA de este sistema es mucho más alto que el de MS, ya que se observa que la totalidad de los valores de MS están por debajo de un 20%. Esto se traduce en las siguientes medias:

Media FA	**14,01%**
Media MS	**1,53%**

Tabla 4-3: Media False Alarm y Missed Speech del SAD.

Teniendo en cuenta que en [11], el rendimiento se evalúa de acuerdo con lo establecido en el apartado 2.3, en donde se penaliza mucho más agresivamente MS que FA, el resultado total es:

Media FA	**14,01%**
Media MS	**1,53%**
DCF	**4,65%**

Tabla 4-4: Rendimiento total del SAD.

Dicho rendimiento parece prometedor ante un conjunto de audio desafiante, como era este caso. Siempre teniendo en cuenta que los resultados presentados en [1] presentaban alrededor de un 7% de error total en un conjunto de audio procedente de radio broadcast.

Por otro lado, se ha considerado relevante, aunque no sea evaluado en [11], el tipo de errores que se han cometido. Esto es, no tener en cuenta sólo el FA o el MS, sino que también se tenga en cuenta las condiciones de ruido que han provocado ese error.

El primer tipo de error que se ha querido estudiar es MS -dado que, además en DCF penaliza más-. Se ha tenido en cuenta la distinta distribución de MS en presencia de locuciones en primer o segundo plano de manera separada, quedando los siguientes histogramas:

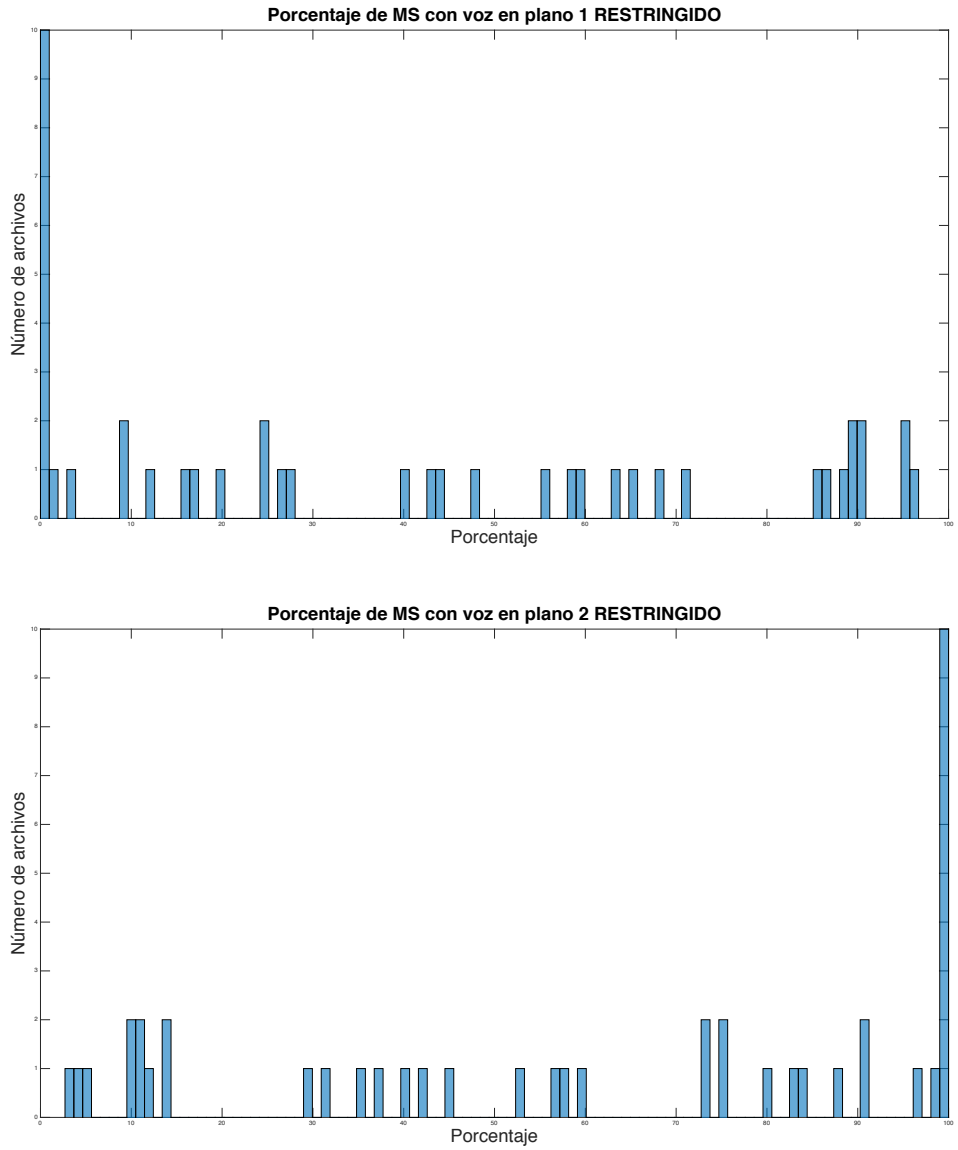

Figura 4-8: Histograma MS para plano 1 (arriba) y MS para plano 2(abajo) por cada fichero del conjunto etiquetado que presentaba voces en el plano 2.

Es importante tener en cuenta que, para que estos resultados tengan más valor se ha restringido el estudio a los ficheros en los que había MS en el plano 2. Si no se hubiese hecho esto quedarían resultados muy desvirtuados debido a la diferencia en número de archivos entre una circunstancia y otra.

En la Figura 4-8 se puede observar que, como podría ser de esperar el valor de MS en voces de segundo plano es mucho más alto que en voces en primer plano. Esto es debido a que la energía de la locución, en comparación con el plano principal, es mucho más bajo y, por tanto, se pasa más por alto.

Media MS plano 1	**39,84%**
Media MS plano 2	**60,15%**

Tabla 4-5: Media Missed Speech para distintos planos SAD.

También se ha estudiado cómo afecta a los valores de MS y FA el hecho de que haya voces de fondo en la grabación. Se podría esperar que esta circunstancia aumentase de manera significativa el valor de FA, ya que se podrían dar las circunstancias para que el sistema confundiese una voz de fondo con una locución en segundo plano, por ejemplo. Sin embargo, este efecto no se ha dado, tal y como se puede ver a continuación:

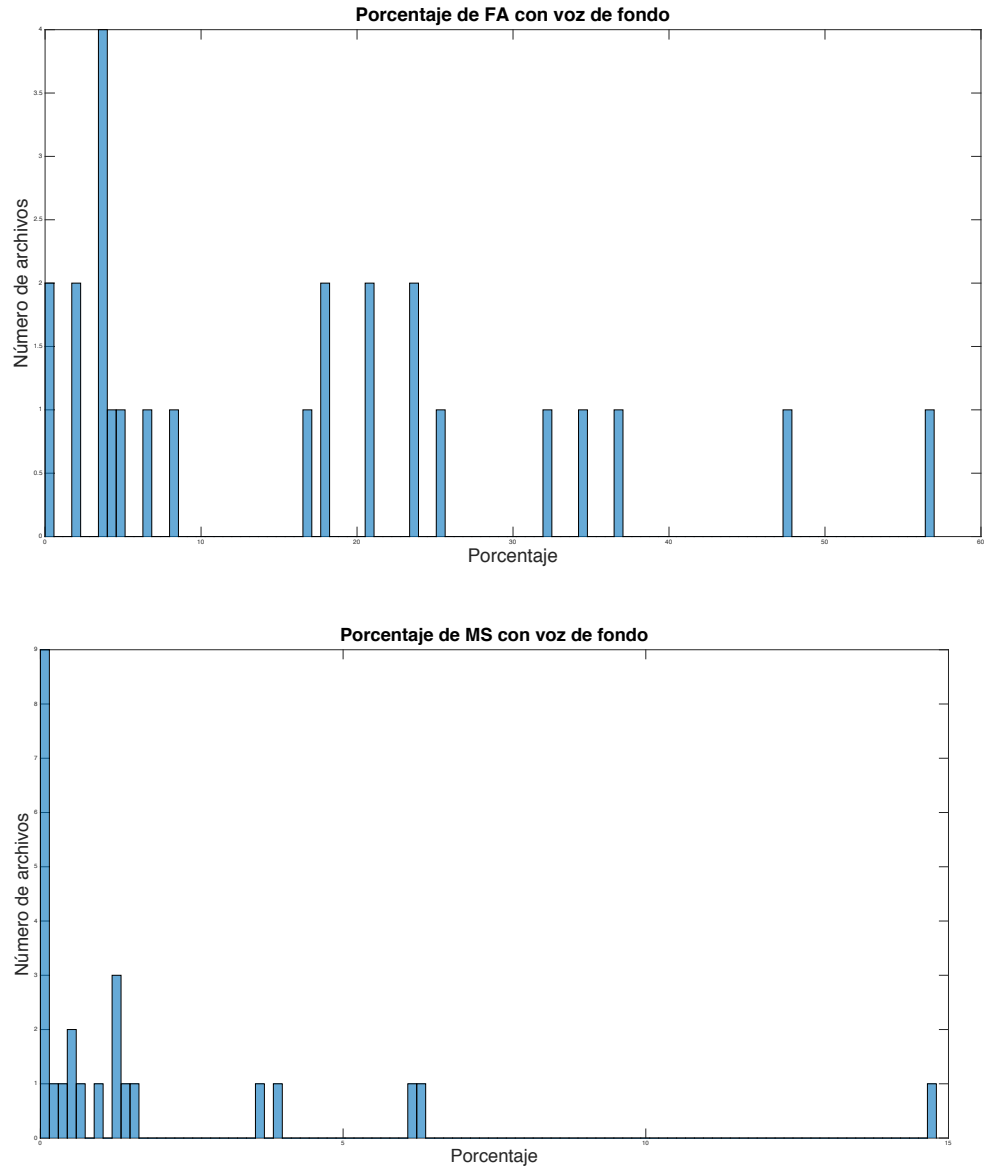

Figura 4-9: Histograma FA con voz fondo (arriba) y MS con voz fondo (abajo) cada fichero del conjunto etiquetado con voces de fondo.

El hecho de que no haya un aumento significativo en el valor de FA -tal y como se podría haber esperado-, se debe al funcionamiento propio de SAD. Si recordamos el apartado 2.1.5 de este TFM, el sistema hace un estudio de las trayectorias de las tramas, descartando aquellas que, bien varían mucho con respecto al anterior, o la energía logarítmica calculada en la etapa de normalización de la ganancia han sacado menos de un 30% del valor de referencia.

Este cálculo de trayectorias hace que las voces de fondo -que no han de ser inteligibles, ya que de serlo pasarían a ser voz en segundo plano-, no pasen dicha etapa de filtrado y, como consecuencia veamos que la FA con voz de fondo no tiene un aumento considerable respecto a la FA en el conjunto de ficheros etiquetados.

4.2.2 Rendimiento SAD + MAD

A continuación, se pasa a estudiar cómo afecta la inclusión del MAD a nuestro sistema de segmentación. Conociendo el corpus, se sabe que no hay una gran presencia de música en el conjunto de audios etiquetados, por lo que, a priori, el sistema MAD no debería afectar en gran medida al rendimiento.

Sin embargo, el hecho de que en este corpus concreto no haya una gran presencia de música, no debe desviarnos de la idea de que el objetivo de todo este TFM consiste en encontrar un sistema que pueda funcionar con la mayor diversidad de audio posible.

Tal y como se pudo ver en [9], en un estudio de la tarea de diarización, se producía un error grande en uno de los conjuntos de audio por la presencia de una melodía que no era detectada correctamente.

De manera análoga a cómo se hizo en 4.2.1, en primer lugar, se comparan los porcentajes de MS y FA por archivo:

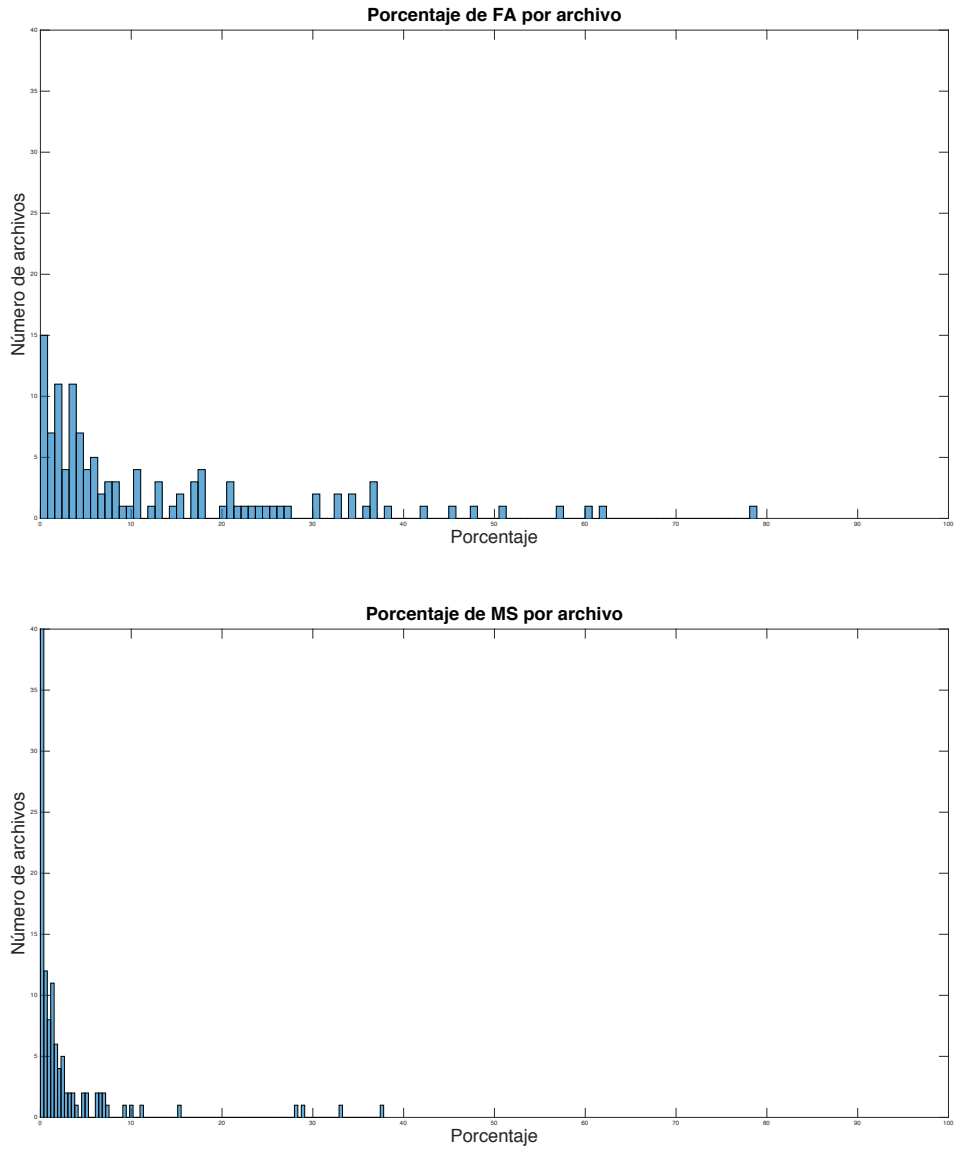

Figura 4-10: Histograma FA (arriba) y MS (abajo) de la combinación de SAD y MAD por fichero del conjunto de audio etiquetado.

Media FA	**13,03%**
Media MS	**2,65%**
DCF	**5,24%**

Tabla 4-6: Rendimiento de la combinación de SAD+MAD.

Comparando los resultados con el sistema SAD individualmente, observamos lo siguiente:

- La media FA mejora ligeramente.
- La media MS empeora casi un punto porcentual.
- En consecuencia, la DCF empeora.

Viendo estos resultados, se puede afirmar que para la evaluación en la que nos estamos centrando, es mejor el sistema compuesto tan solo por el SAD. Sin embargo, es importante tener en cuenta que, esa mejora en la media FA puede ser interesante para otras aplicaciones que no penalicen tanto esa medida de MS.

Además, se debe tener en cuenta la universalidad de un sistema resultante con ambos sistemas. Esto es, que el sistema resultante pueda utilizarse para cualquier corpus disponible sin necesidad de adaptarlo. Así pues, un sistema que incluya un MAD podrá mantener su efectividad en un corpus con una gran presencia de música y melodías.

4.2.3 Rendimiento de la DNN

El primer hecho para tener en cuenta al evaluar el rendimiento de una DNN es que el mismo estará fuertemente influenciado por el entrenamiento que se haya realizado. En nuestro caso se ha decidido, en colaboración con el desarrollador de [2], que el conjunto de datos de entrenamiento incluiría el 85% de los conjuntos de datos VAST y BABEL. De esta manera, mantenemos un 15% de los mismos para la etapa de validación dentro de la DNN -este conjunto no se utiliza para obtener los pesos de las conexiones entre neuronas-.

En la DNN se va a distinguir entre varios resultados posibles. Tal y como se ha comentado en 2.2, el resultado que nos devuelve la DNN es filtrado con el objetivo de eliminar valores intermedios que se traducirían, en una locución, en transiciones no naturales dentro de la misma.

4.2.3.1 Sin filtrado

El primer caso que se va a evaluar es el de que no se hace tal filtrado después de la ejecución de la DNN. De manera análoga a los otros dos sistemas, los porcentajes de MS y FA quedan distribuidos de la siguiente forma:

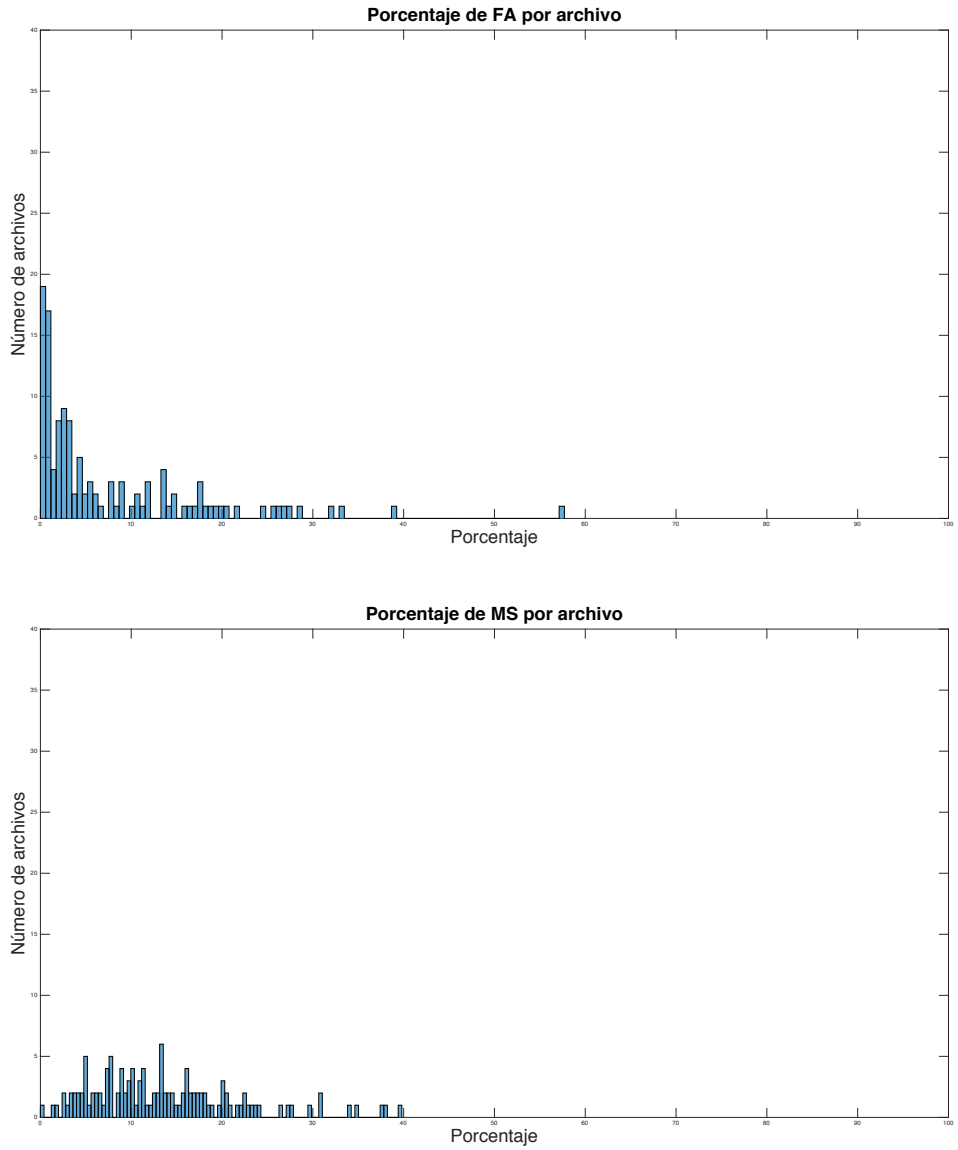

Figura 4-11: Histograma FA (arriba) y MS (abajo) para la DNN sin filtrado por cada fichero del conjunto etiquetado.

Media FA	**7,62%**
Media MS	**13,60%**
DCF	**12,11%**

Tabla 4-7: Rendimiento DNN sin aplicar filtrado.

En este caso se puede observar una mejora importante en FA con respecto a los sistemas basados en correlación, SAD y SAD+MAD. Sin embargo, esa misma mejora se obtiene como empeoramiento en MS y, dado que, como ya se ha dicho, el MS es más importante para el cálculo de la DCF, este sistema, de por sí, no presenta alternativa a SAD.

4.2.3.2 Filtrado de una décima de segundo

En segundo lugar, se pasa a realizar el filtrado de silencios intermedios en la locución de una décima de segundo.

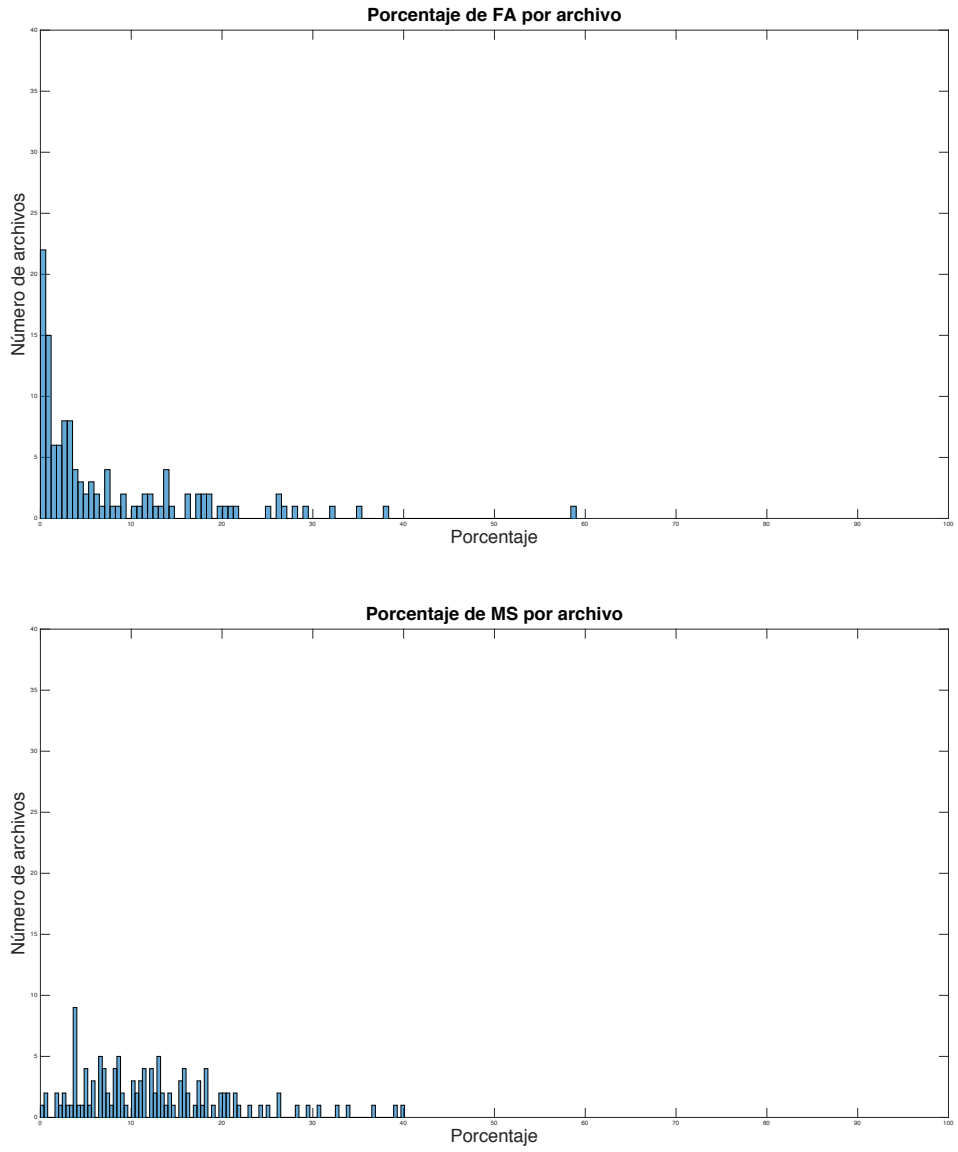

Figura 4-12: Histograma FA (arriba) y MS (abajo) para la DNN añadiendo un filtrado de una décima de segundo.

Media FA	**7,62%**
Media MS	**12,60%**
DCF	**11,35%**

Tabla 4-8: Rendimiento DNN filtrado a la décima de segundo.

Tal y como era de esperar tras haber realizado el etiquetado, la introducción de un filtrado de los resultados con el objetivo de eliminar del sistema esas transiciones no naturales entre voz y no voz, mejora, con sólo introducir una décima de segundo de filtrado, un punto porcentual la media de MS.

Por otra parte, sabemos que la precisión del etiquetado está realizada con, aproximadamente, una tolerancia de medio segundo para la asignación de silencio a una trama concreta. Este hecho hace que sea interesante probar los resultados con la mitad de la tolerancia, esto es, filtrando un cuarto de segundo tras la DNN.

4.2.3.3 Filtrado de un cuarto de segundo

En este caso, se podrá esperar una ligera mejora respecto al anterior caso, teniendo en cuenta dos factores:

1. El filtrado de una décima de segundo no representa la tolerancia que se tomó como base en la fase de etiquetado.
2. El efecto en la mejora de resultados no será tan grande como de pasar de no haber filtrado a que lo haya, aunque sea pequeño, en parte por la presencia del collar en la evaluación.
3. Según se van mejorando los resultados, mejorar es cada vez más costoso.

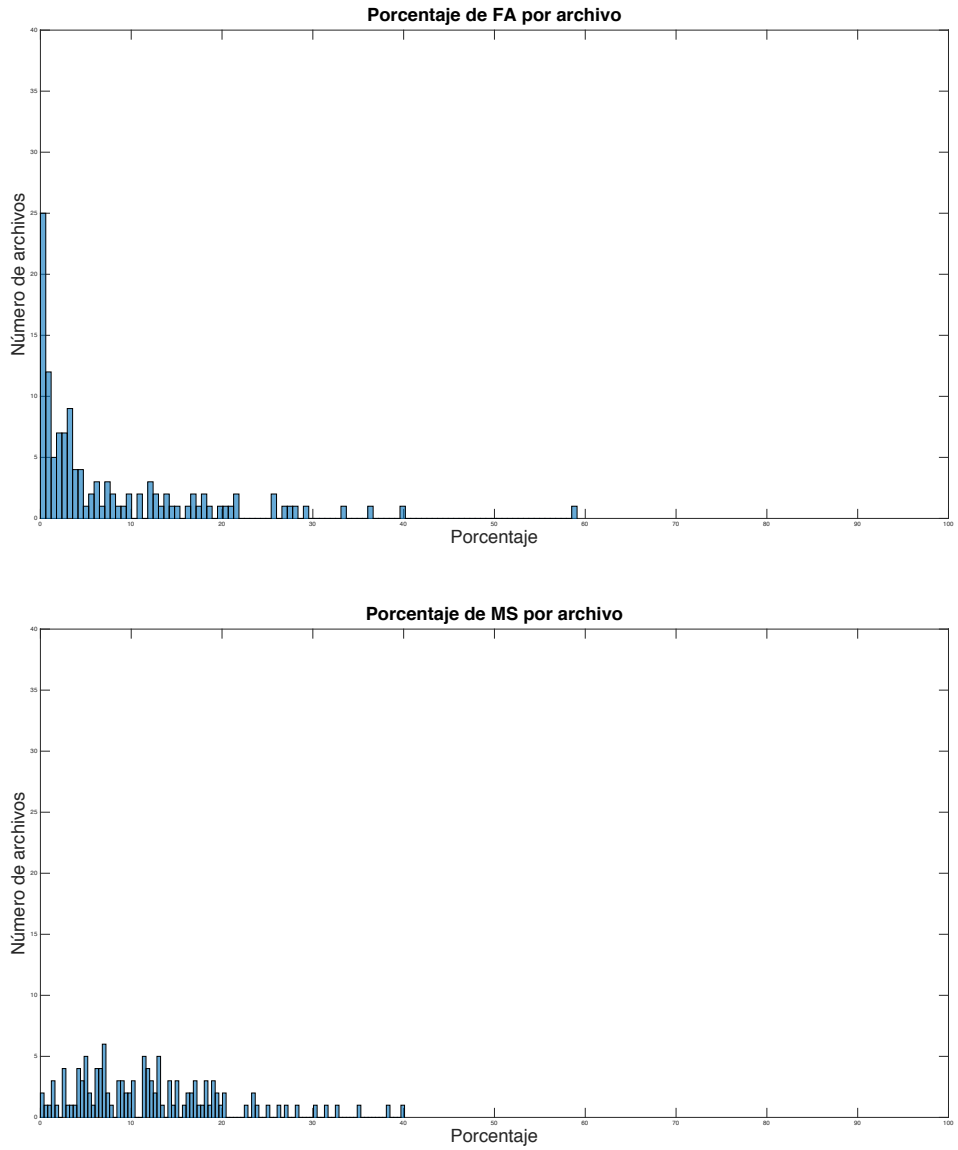

Figura 4-13: Histograma FA (arriba) y MS (abajo) para la DNN añadiendo un filtrado de un cuarto de segundo.

Media FA	**7,69%**
Media MS	**12,13%**
DCF	**11,02%**

Tabla 4-9: Rendimiento DNN filtrado a un cuarto de segundo.

Tal y como era de esperar, se ha mejorado ligeramente los resultados con respecto al sistema anterior, aunque ya se aprecia un ligero empeoramiento en FA. Sin embargo, el

rendimiento total de este sistema sigue estando muy lejos del SAD estudiado para esta evaluación.

4.2.3.4 Filtrado de medio segundo

Por último, se prueba con un filtrado de medio segundo. Este es el valor más alto que tiene sentido utilizar, ya que en la etapa de etiquetado sí que se han marcado silencios o locuciones de medio segundo, o más, de duración.

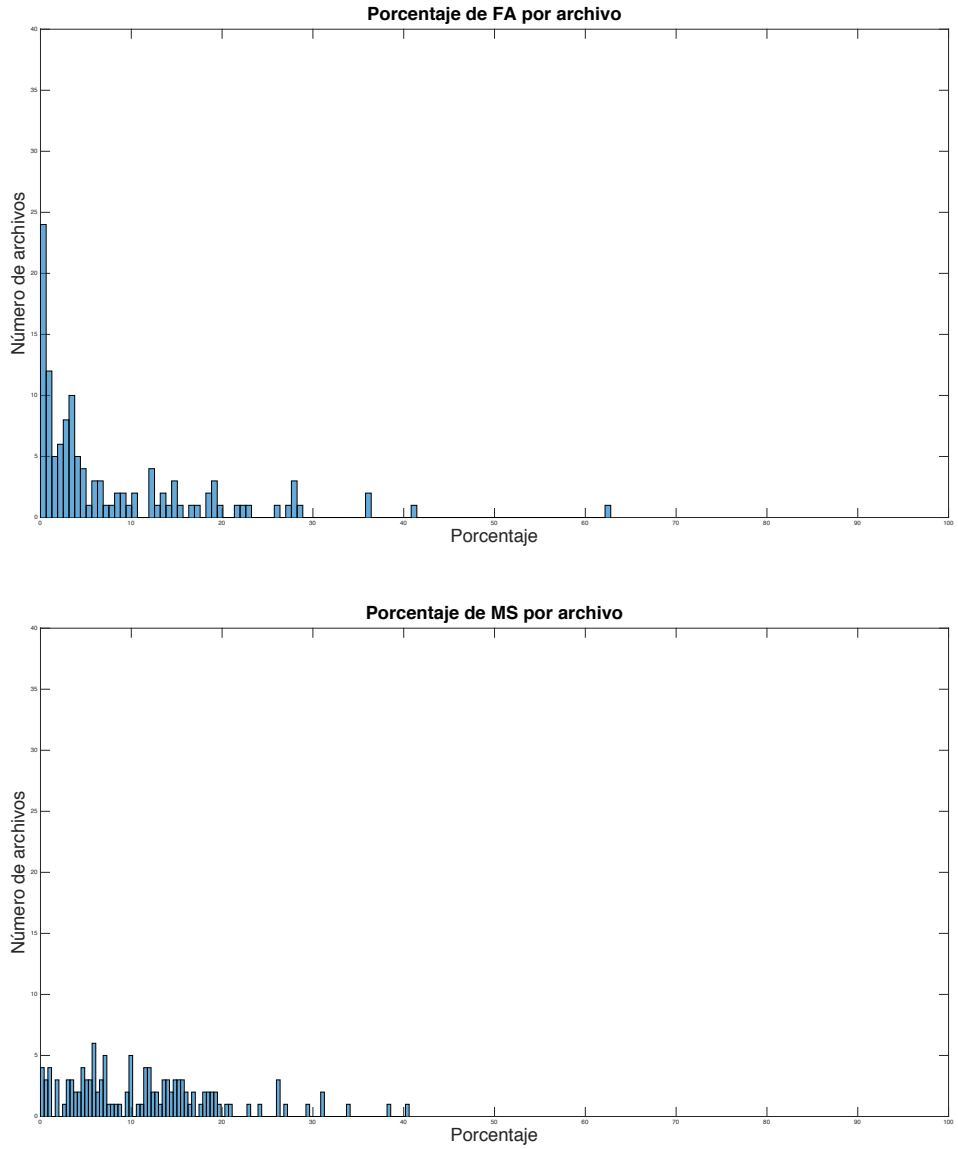

Figura 4-14: Histograma FA (arriba) y MS (abajo) para la DNN añadiendo un filtrado de medio segundo.

Media FA	**7,91%**
Media MS	**11,40%**
DCF	**10,53%**

Tabla 4-10: Rendimiento DNN filtrado a medio de segundo.

Tal y como era de esperar, los mejores resultados se obtienen en el momento en el que somos más laxos con el filtrado que hacemos a la DNN. Sin embargo, se ha de tener en cuenta que, ya en la última prueba, se ha aumentado el valor de FA, por lo que, aunque obviásemos el hecho de que no tiene sentido filtrar más de medio segundo por el etiquetado realizado, se empezaría a entrar en un compromiso en el que la mejora de FA implicaría un empeoramiento de MS.

4.2.4 Fusión de sistemas

En este punto se van a probar distintas fusiones de cada uno de los sistemas con el objetivo de optimizar el sistema final.

4.2.4.1 Fusión I

La primera alternativa de fusión va a tomar como sistemas de referencia el SAD y la DNN con el filtrado de medio segundo, ya que estos han sido los que mejores resultados han obtenido.

Estudiando las fortalezas de cada sistema, podemos observar que el SAD presentaba unos valores de MS tremendamente ajustados, mientras que la gran fortaleza del sistema basado en DNN era una mejora en FA.

Por esta razón, la primera fusión que se plantea es un sistema tal que, si el SAD determina que ese segmento es de *speech*, la fusión decidirá *speech*. Sin embargo, si el SAD determina *no-speech*, se pasará a mirar el resultado de la DNN. En caso de que la DNN reafirme el *no-speech* se acabará, pero si la DNN dice *speech* con una confianza mayor que un umbral se cambiará el resultado del SAD.

Media FA	**14,24%**
Media MS	**1,31%**
DCF	**4,54%**

Tabla 4-11: Rendimiento Fusión I con umbral = 0,8.

Media FA	**14,73%**
Media MS	**0,98%**
DCF	**4,42%**

Tabla 4-12: Rendimiento Fusión I con umbral = 0,6.

Media FA	**15,15%**
Media MS	**0,80%**
DCF	**4,39%**

Tabla 4-13: Rendimiento Fusión I con umbral = 0,5.

No se puede bajar el umbral más de 0,5, ya que eso implicaría que la DNN ha decidido *no-speech.* Esta circunstancia hace que esta fusión de sistemas en concreto se corresponda con una decisión por unanimidad, esto es, si ambos sistemas dicen que hay *speech*, la salida final será *speech* y bastará con que uno solo de los sistemas no refrende esa decisión para dar salida *no-speech*.

Dibujando el resultado de este sistema en un *scatter plot* podemos ver:

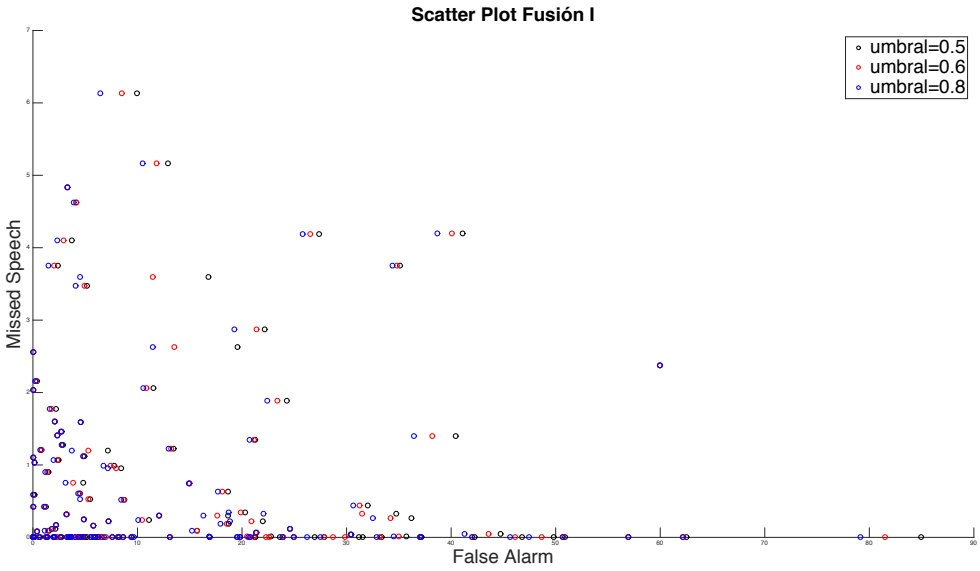

Figura 4-15: Scatter plot para el sistema Fusión I. Cada punto representa el valor medio de FA (abcisas) y MS (ordenadas) de un fichero de los etiquetados en la primera fase del TFM.

En primer lugar, hay que tener en cuenta la diferencia en los ejes del *scatter plot*, se ha dado más resolución al eje MS -llegando tan solo a 7%-, con el objetivo de poder distinguir lo que ocurre en las partes más bajas de MS. Recordemos que MS es el parámetro más importante para nuestra métrica de rendimiento.

Podemos observar una gran concentración de puntos en el mismo eje de Falsa Alarma -MS aproximadamente cero-, y, en general, se ven puntos muy bajos en el eje de ordenadas. Esto implica que, dejando la media aparte, la mayoría de los puntos tendrán unos muy

buenos resultados, mientras que unos pocos archivos harán subir la media de manera considerable.

En cuanto al cambio que supone en el *scatter plot* la variación del umbral, vemos que, en su mayoría, los puntos se van desplazando de derecha a izquierda según vamos aumentando el umbral, hecho que explica la mejora de False Alarm observada con umbral=0,8.

4.2.4.2 Fusión II

Partiendo de los mismos dos sistemas que en Fusión I -sistema SAD y DNN con filtrado de medio segundo-, pasamos a probar otra alternativa de decisión entre ambos sistemas. En este caso probaremos una decisión basada en que con que uno de los sistemas clasifique un segmento como voz, será suficiente para que la salida del sistema sea clasificada como tal.

Media FA	**14,76%**
Media MS	**0,92%**
DCF	**4,38%**

Tabla 4-14: Rendimiento Fusión II.

En este caso, respecto al caso anterior, podemos observar una ligera degradación en el resultado de MS a cambio de una mejora importante de FA, haciendo que el global de DCF caiga un 0,01%. Sin embargo, dicha mejora global en DCF no justifica utilizarlo por encima del sistema Fusión I. La diferencia es tan pequeña que puede depender exclusivamente del conjunto de audio etiquetado y, sigue siendo preferible mejorar en MS a FA.

En este caso el *scatter plot* resultante es:

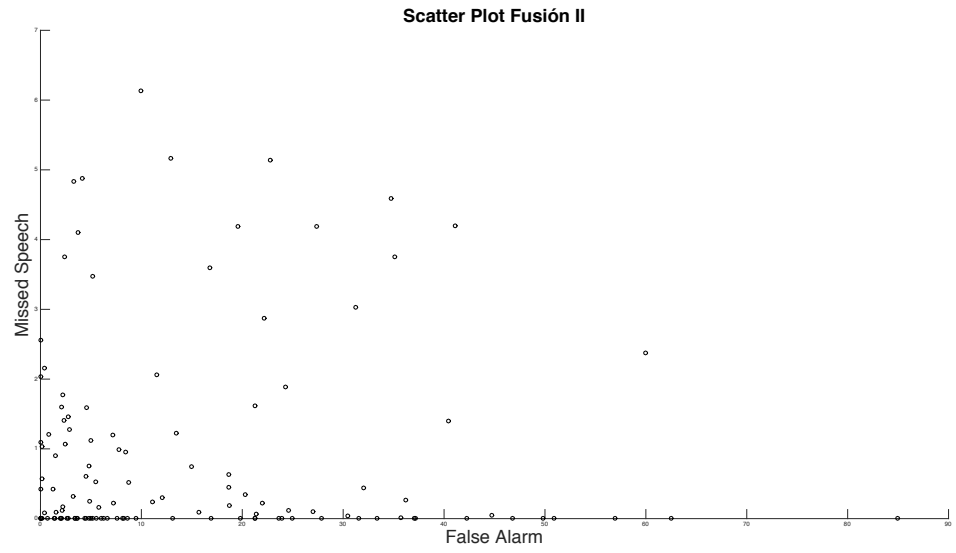

Figura 4-16: Scatter plot para el sistema Fusión II. Cada punto representa el valor medio de FA (abcisas) y MS (ordenadas) de un fichero de los etiquetados en la primera fase del TFM.

Como era de esperar tras los primeros resultados, el *scatter plot* para este sistema es similar al anterior. El único desplazamiento resultante es aquel que hace desviar las medias ligeramente, pero no hay *outliers* que no se presentasen en el caso anterior.

4.2.4.3 Fusión III

En esta prueba, se utilizará el filtrado tras la DNN de un cuarto de segundo, con el SAD y bastará con que uno de los sistemas catalogue un segmento como *speech* para que esa sea la decisión final. Se ha decidido estudiar este caso, por si una pérdida de eficiencia por efecto del filtrado podía ser compensada con el SAD y, en global, se podía mejorar rendimiento.

Media FA	**15,09%**
Media MS	**0,89%**
DCF	**4,44%**

Tabla 4-15: Rendimiento Fusión III.

En este caso, respecto a su homólogo Fusión II, se observa una mejora de MS desde 0,92 a 0,89. Esta mejora queda compensada en el cálculo de DCF por la degradación sufrida en FA de más de un 0,3%. Por esta razón este sistema queda desechado frente a Fusión II.

El efecto que ha aportado el filtrado de los resultados de la DNN consiste en que algunos segmentos que, cuando ese filtrado era de medio segundo, eran clasificados como voz por la DNN, ahora pasarán a ser no voz, debido a la menor duración del filtrado.

4.2.4.4 Fusión IV

Pese a que los resultados de la integración de SAD y MAD no han mejorado al sistema SAD, se decide probar la integración de ambos sistemas con la DNN para el resultado final.

La razón de esta prueba radica en que uno de los objetivos principales de este TFM es el de proporcionar a su término un sistema que, teniendo buen rendimiento, se pueda utilizar en toda clase de bases de datos de audio.

Este corpus en concreto no presenta cantidades suficientes de música como para que sea interesante su utilización en un primer lugar, sin embargo, la degradación en el rendimiento del SAD puede ser compensada en otro corpus con una presencia de música mayor.

	Umbral = 0,5	*Umbral = 0,6*	*Umbral = 0,8*
Media FA	14,76	14,25	13,52
Media MS	0,92	1,17	1,78
DCF	4,38	4,44	4,71

Tabla 4-16: Rendimiento del sistema resultado Fusión IV con distintos umbrales de decisión.

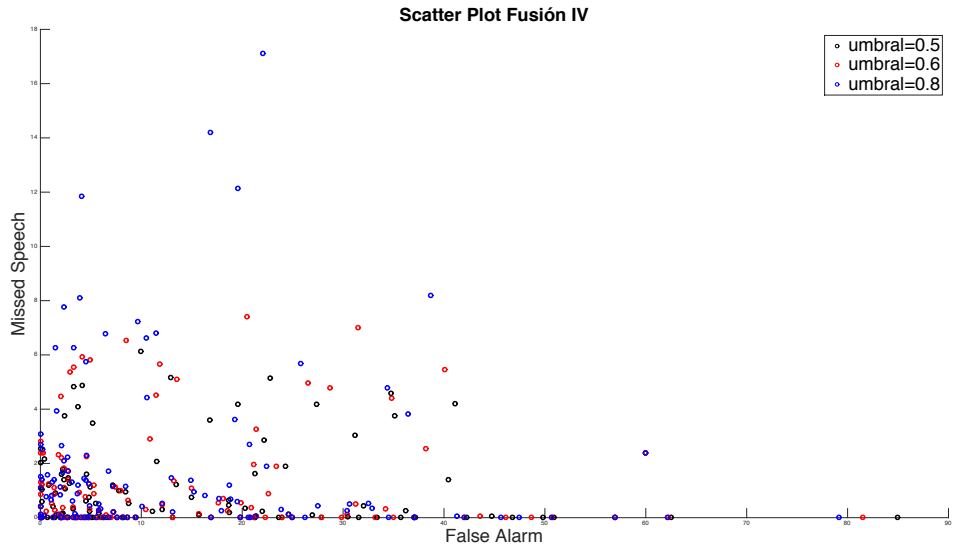

Figura 4-17: Scatter plot para el sistema Fusión IV, superpuestos los resultados para los distintos valores de umbral. Cada punto representa el valor medio de FA (abcisas) y MS (ordenadas) de un fichero de los etiquetados en la primera fase del TFM.

En la Figura 4-16 se puede observar claramente cómo, conforme se disminuye el umbral hasta 0,5 se eliminan los *outliers* superiores que causaban una subida generalizada de la media de MS en el sistema.

Finalmente, hemos dado con un sistema con un rendimiento que mejora nuestro valor inicial marcado con SAD -de 4,65% de DCF a 4,38%-. Pero no solo es importante este hecho, sino que también, se ha conseguido universalizar al máximo el sistema para todo tipo de corpus de entrada con el que queramos utilizarlo, ya que combina las fuerzas del detector SAD, el detector MAD y el sistema basado en DNN.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

A la hora de mirar las conclusiones es importante tener en cuenta los objetivos propuestos en el capítulo 1. El objetivo principal que se planteaba era el de hacer una medida del rendimiento de distintos sistemas presentados a la evaluación NIST OpenSAT. De manera secundaria, también se ha realizado una parte de trabajo importante de cara a futuro, como es el inicio en la creación de un nuevo corpus de audio, con un etiquetado con alto nivel de detalle.

Finalmente se ha logrado el objetivo fundamental de este TFM, conseguir un sistema de segmentación de audio que, no solo consiga buenos resultados en la tarea de segmentación de audio en corpus desafiantes, sino que también, se ha planteado un sistema final que, a priori, será robusto a cualquier otro corpus con el que se quiera probar.

Así pues, a lo largo de este TFM se ha realizado lo siguiente:

- Se ha realizado un estudio de los sistemas presentados a NIST OpenSAT, estos son: un Speech Activity Detector, Music Activity Detector y una Deep Neural Network.

- Se ha reentrenado el sistema de la DNN, ya que el entrenamiento existente utilizaba nuestro conjunto de audio etiquetado como parte del corpus de entrenamiento.

- Se ha creado un corpus de audio a partir de la Base de Datos VAST proporcionada para la evaluación NIST OpenSAT.

- Se ha desarrollado la algoritmia necesaria para las medidas de rendimiento, todas ellas calculadas a través de Matlab.

- Se ha hecho un estudio de los resultados de los sistemas, cambiando parámetros de funcionamiento cuando era posible.

- Se ha llegado a proponer un sistema final con buenos resultados.

5.2 Trabajo futuro

- Un paso interesante de cara a futuros trabajos en esta materia sería la ampliación del nuevo etiquetado a la totalidad de VAST y la introducción del mismo en el conjunto BABEL. En primer lugar, para poder tener unos conjuntos aún más diversos y extensos en el caso de VAST y, por otro lado, observar el comportamiento del sistema en un conjunto de audio tan controlado como es BABEL.

- Probar este nuevo sistema con una mayor cantidad de datos para el entrenamiento de la DNN. Si es necesario en un nuevo corpus.

- Una línea importante que se podría también es la de incorporar datos de la Base de Datos "Audias-ATVS-Radio", desarrollada en [1], [7], [8] y [9]. Estos niveles de etiquetado, siendo distintos, son compatibles a nivel de voz con el etiquetado realizado en este TFM. Y sería interesante comprobar el rendimiento de este sistema en un entorno controlado como es el audio *broadcast*.

- Hemos visto también que el sistema basado únicamente en DNN obtiene mejores resultados, si ponderamos de igual manera MS y FA, que los demás sistemas. Sin embargo, la mejora en FA se ve malograda por una bajada en la medida MS, lo que hace que, usando la métrica NIST, se empeore el resultado total. Es por esto, que una modificación en la estructura de la propia DNN, con el fin de mejorar los resultados de MS podría ser interesante.

Referencias

[1] B. García Naranjo y J. González Rodríguez, “Segmentación de Audio en Audio Broadcast”, Madrid: Universidad Autónoma de Madrid, 2016.

[2] Lozano-Diez A, Zazo R, Toledano DT, Gonzalez-Rodriguez J (2017) An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition. PLoS ONE 12(8): e0182580.

[3] Álvaro Escudero-Barrero, Alicia Lozano-Diez, Ruben Zazo, Javier Franco-Pedroso, Doroteo T. Toledano, Joaquín González Rodríguez, System Description para NIST OpenSAT 2017. “Speech Activity Detection Combining DNN-Trained and Rule-Based Voice and Music Detectors at NIST OpenSAT 2017”, 2017. Disponible tras petición a Audias-ATVS Group.

[4] Ruben Zazo, Tara N. Sainath, Gabor Simko, Carolina Parada, “Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection”, Interspeech 2016, pp. 3668-3672.

[5] Z. XL and D. Wang, “Boosted Deep Neural Networks and Multiresolution Cochleagram Features for Voice Activity Detection”, Interspeech, 2014, pp. 1534-1538.

[6] Aythami Morales Moreno, Transparencias para la asignatura del Máster Universitario en Ingeniería de Telecomunicación, “Procesado Avanzado de Señal para Multimedia”, Tema 1.5 Redes Neuronales, 2016.

[7] M. P. Fernández Gallego y D. Torre Toledano, “Mejoras de un Sistema de Búsquedas de Voz y Aplicación a Detección de Menciones en Medios de Comunicación”. Madrid: Universidad Autónoma de Madrid, 2016.

[8] Á. Escudero Barrero y J. González Rodríguez, “Búsqueda Eficiente de Audio Pregrabado en Audio Broadcast”, Madrid: Universidad Autónoma de Madrid, 2016.

[9] G. Soriano Morancho y J. González Rodríguez, “Diarización de Locutores en Audio Broadcast”, Madrid: Universidad Autónoma de Madrid, 2016.

[10] Z. H.-J. J. H. Lu Lie, “Content analysis for audio classification and segmentation” de IEEE Transactions on Speech and Audio Processing, vol. 10, Beijing, IEEE, 2002, pp. 504-516.

[11] NIST OpenSAT Pilot Evaluation Plan, DRAFT version 1.1, 2017. Disponible en: https://www.nist.gov/file/364071

[12] D. de Benito Gorrón y J. González Rodríguez, “Detección de Música en Contenidos Multimedia mediante Ritmo y Armonía”, Madrid: Universidad Autónoma de Madrid, 2017.

[13] M.H. Moattar and M. M. Homayounpour, “A Simple but Efficient Real-Time Voice Activity Detection Algorithm”, 17th European Signal Processing Conference (EUSIPCO 2009) pp. 2549-2553.

Glosario

TFM	Trabajo Fin de Máster
NIST	National Institute of Standards and Technology
VAD	Voice Activity Detection
MS	Missed Speech
FA	False Alarm
DCF	Detection Cost Function
SAD	Speech Activity Detector
MAD	Music Activity Detector
DNN	Deep Neural Network
DEV	Development
EVAL	Evaluation
MFCC	Mel Frequency Cepstral Coefficients